Functional Subspace Watermarking for Large Language Models

Zikang Ding [* 1 2]   Junhao Li [* 3]   Suling Wu [1]   Junchi Yao [1 2]   Hongbo Liu [1]   Lijie Hu [2]

Abstract

Model watermarking utilizes internal representations to protect the ownership of large language models (LLMs). However, these features inevitably undergo complex distortions during realistic model modifications such as fine-tuning, quantization, or knowledge distillation, making reliable extraction extremely challenging. Despite extensive research on model-side watermarking, existing methods still lack sufficient robustness against parameter-level perturbations. To address this gap, we propose **`Functional Subspace Watermarking (FSW)`**, a framework that anchors ownership signals into a low-dimensional functional backbone. Specifically, we first solve a generalized eigenvalue problem to extract a stable functional subspace for watermark injection, while introducing an adaptive spectral truncation strategy to achieve an optimal balance between robustness and model utility. Furthermore, a vector consistency constraint is incorporated to ensure that watermark injection does not compromise the original semantic performance. Extensive experiments across various LLM architectures and datasets demonstrate that our method achieves superior detection accuracy and statistical verifiability under multiple model attacks, maintaining robustness that outperforms existing state-of-the-art (SOTA) methods.

Figure 1. Conceptual comparison of watermark robustness under common model-side attacks. **Top (Previous):** Conventional internal watermarks are vulnerable to post-hoc modifications like fine-tuning, quantization, pruning, and distillation, which often result in signal erasure. **Bottom (FSW):** Our proposed framework anchors ownership signals into a stable functional backbone, ensuring that the watermark remains detectable even after significant parameter-level perturbations.

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 signal (As shown in Figure 1). Model internal watermarking for LLMs remains relatively under explored, which is still a central challenge.

To address this gap, we propose **`FSW`**, which embeds ownership signals into the functional backbone of a model. We identify a low-dimensional latent subspace that is critical to

[1]University of Electronic Science and Technology of China, Chengdu, China [2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates [3]South China University of Technology, Guangzhou, China. Correspondence to: Lijie Hu <lijie.hu@mbzuai.ac.ae>.

task performance and remains stable under a broad class of compression-related transformations, enabling robustness to common model modifications. Based on this subspace, we formulate a generalized eigenvalue framework that explicitly decouples task-critical sensitivity from compression-induced variance, allowing the extraction of a stable watermark carrier. Furthermore, we introduce a spectral truncation strategy that selects a specific range to achieve an optimal balance between robustness and model utility. Experimental results show that this targeted selection maintains high watermark verifiability while strictly preserving the original model performance under various model modifications, including fine-tuning, quantization, and knowledge distillation.

This paper makes the following key contributions:

1. We propose the **FSW** framework, which embeds watermarks by locating a functional backbone in the model that is both critical and stable. This method uses a generalized eigenvalue problem to extract carrier directions, effectively defending against various compression transformations and distillation attacks.

2. We design an orthogonal-key-based multi-bit watermark verification scheme, which transforms the ownership verification process into a rigorous statistical test. Without affecting the model's original generation performance, this scheme achieves reliable watermark injection and accurate verification.

3. We conduct extensive experiments on multiple model architectures and datasets to validate the robustness of **FSW** against fine-tuning, quantization, and knowledge distillation. The results demonstrate that, while maintaining model performance, the framework can complete ownership verification at an extremely low false positive rate.

2. Related Work

2.1. Content Watermarking

Content watermarking embeds statistically detectable signals into generated texts to enable provenance tracing without requiring access to model parameters. Many approaches operate at generation time by steering token selection with a secret key so that the resulting text exhibits a measurable signature. A representative instance is the greenlist-based design of Kirchenbauer et al. (2023), and subsequent systems improve robustness, payload, and sampling efficiency across broader settings (Zhang et al., 2024; Mao et al., 2025; Niess & Kern, 2025). Moving toward deployment, Dathathri et al. (2024) demonstrate watermarking at practical scale

for identifying LLM outputs, showing that such detectors can be integrated into real inference pipelines.

Recent work further explores robustness against post generation edits and stronger verification. Dabiriaghdam & Wang (2025) is notable for shifting the carrier from token statistics to sentence level semantic similarity, improving robustness to paraphrasing while remaining compatible with API-only access. Another direction targets stronger payload under paraphrasing by explicitly leveraging LLM-based paraphrasers to encode and recover multi-bit messages (Xu et al., 2025). Orthogonal to these, ensemble-style schemes combine multiple watermark signals to improve detectability and resilience across diverse attack patterns (Niess & Kern, 2025), while in-context watermarking explores watermark injection and detection through prompt-time mechanisms rather than modifying the underlying model (Liu et al., 2025). On the theory side, distribution adaptive frameworks aim to ground watermark design in statistical guarantees under varying output distributions (He et al., 2024).

2.2. Model Watermarking

Model watermarking aims to embed ownership information into model parameters or internal representations so that suspect models can be verified in ownership disputes, even after common post-hoc modifications. Classic approaches in deep neural networks include trigger-set and backdoor based ownership proofs (Adi et al., 2018), end-to-end watermark embedding frameworks (Darvish Rouhani et al., 2019), and covert, robust white-box watermarking designs (Wang & Kerschbaum, 2021).

For LLMs, internal watermarking remains relatively under-explored. Existing efforts include embedding signals through weight quantization (Li et al., 2023), leveraging "radioactive" training traces for downstream attribution (Sander et al., 2024), and SEAL, which anchoring ownership signals in latent subspaces (Dai et al., 2025). Related black-box ownership verification via fingerprinting has also been studied under model merging, for example MERGEPRINT (Yamabe et al., 2025). Despite progress, robustness under strong extraction, especially knowledge distillation, remains a key challenge. This leaves an important gap for model-side watermarking methods designed explicitly for distillation settings. To address this gap, we propose **FSW**, which embeds ownership information in functional subspaces that are important for the task and remain stable under compression and distillation, improving verifiability under distillation threats.

3. Threat Model

We consider a model ownership protection scenario where a model owner releases a watermarked LLM and subsequently
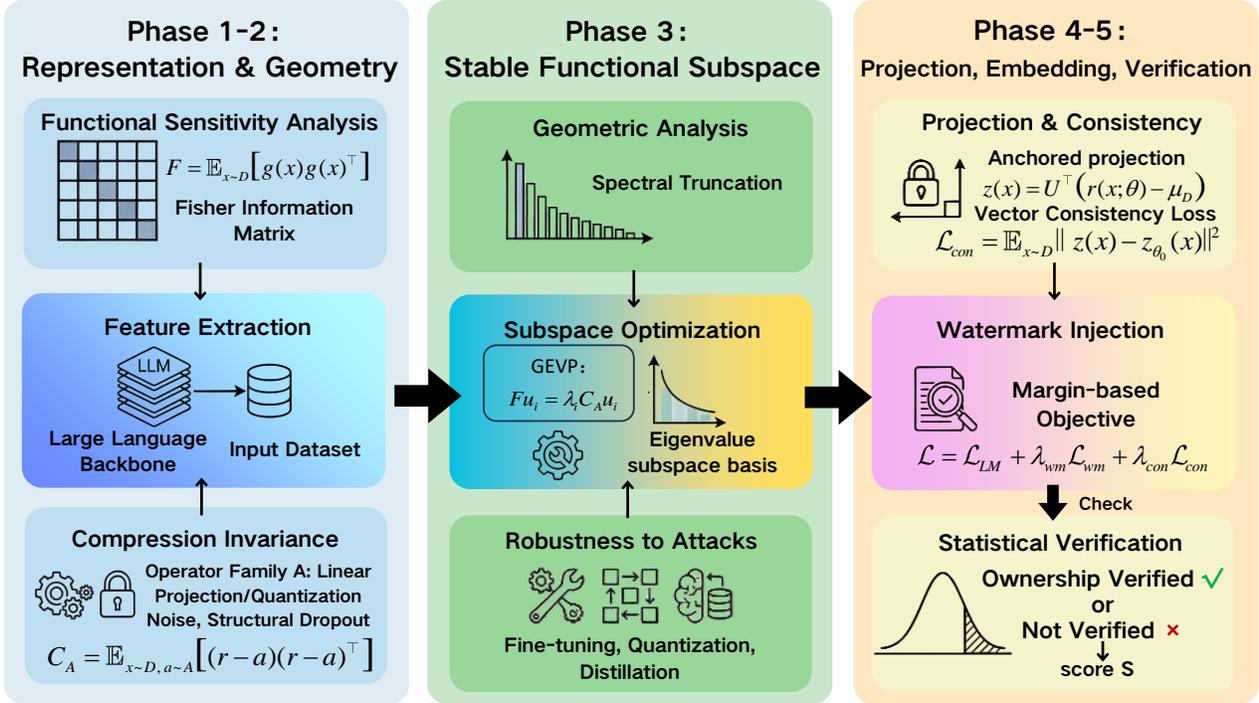
Figure 2. Overall framework of FSW.

identifies a suspect model of disputed provenance. The adversary's objective is to remove or invalidate the embedded signature while maintaining the model's practical utility for downstream tasks (Wang & Kerschbaum, 2021; Xu et al., 2024).

Adversary Capabilities and Modifications. Following the taxonomy of model lineage auditing (Shao et al., 2025), we assume a strong adversary with white-box access to the model parameters (Pasquini et al., 2025). The adversary can perform various post-hoc modifications to evade detection, including supervised fine-tuning (SFT), parameter-efficient adaptation (e.g., LoRA), and model compression techniques such as weight pruning and quantization (Xu et al., 2024; Yan et al., 2023). Crucially, the adversary is permitted to optimize the modified model to restore task-level performance, ensuring it remains competitive in terms of perplexity and downstream accuracy (Xu et al., 2024).

Functional Preservation Constraint. Consistent with the requirements for robust internal watermarking (Yao et al., 2024), we restrict the adversary to attacks that preserve the model's internal functional backbone representations. This constraint reflects realistic deployment settings where adversaries prioritize the structural integrity and reliability of the original model over re-engineering its internal logic from scratch. Accordingly, our guarantees do not extend to fully unconstrained distillation where the student model is free to arbitrarily reparameterize its latent space (Shao et al., 2025; Xu et al., 2024).

Verification Protocol. We assume the defender has white-box access to the suspect model for ownership verification (Wang & Kerschbaum, 2021). This allows for the inspection of model parameters or intermediate activations to establish proof of ownership (Li et al., 2025). By analyzing representations within the identified functional subspace, the defender can recover embedded signatures even after significant parameter-level perturbations or structural obfuscations (Yan et al., 2023).

4. Methodology

We propose `FSW`, which embeds ownership signatures into a stable *functional backbone* by solving a Generalized Eigenvalue Problem (GEVP). This backbone is optimized for task-criticality and compression resilience to ensure watermark persistence during distillation or quantization. The implementation follows a structured four-phase pipeline:

4.1. Phase 1: Representation Extraction

Let f_θ denote an autoregressive language model with L layers and hidden dimension d. Given an input sequence $x = (x_1, \ldots, x_T)$, we extract the representation from a

fixed intermediate layer ℓ:

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

where $H_T^{(\ell)}$ is the hidden state of the last token. This state serves as the information bottleneck for next-token prediction, making it the optimal carrier for semantic-level watermarking.

4.2. Phase 2: Geometry Analysis

To construct a stable watermark injection subspace, we analyzed the characteristics of the subspace from two aspects:

Functional Sensitivity. We quantify the task relevance of representation directions using the Fisher information matrix F, estimated on a calibration set drawn from the target distribution $\mathcal{D}$:

$$F = \mathbb{E}_{x \sim \mathcal{D}}\left[g(x)g(x)^\top\right] \in \mathbb{R}^{d \times d}, \tag{2}$$

where $g(x) = \nabla_r L_{\text{LM}}(x;\theta)$. Directions with large quadratic forms $u^\top F u$ encode the model's core knowledge. Perturbing these directions significantly alters the output distribution, making them inherently resistant to pruning-based removal (Yan et al., 2023).

Invariance to Compression Operators. Rather than overfitting to a specific student architecture, we model knowledge distillation as a process of information compression. We define a family of Compression Operators $\mathcal{A}$ that approximate shared invariants, including:

- *Linear Projection:* $a(r) = W_{\text{low}}r$, 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.

We define the transformation-invariant matrix $C_{\mathcal{A}}$ as the expected variance caused by these operators:

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

Minimizing $u^\top C_{\mathcal{A}} u$ encourages the watermark to reside in directions that remain stable under compression-equivalent transformations (Yao et al., 2024).

4.3. Phase 3: Subspace Construction & Optimization

We seek a subspace $U^\star \in \mathbb{R}^{d \times k}$ that maximizes the signal-to-noise ratio between functional sensitivity and compression-induced variance. This objective is formulated as the GEVP:

$$Fu_i = \lambda_i C_{\mathcal{A}} u_i, \tag{4}$$

where the generalized eigenvalue λ_i represents the robustness score of the i-th direction.

Spectral Truncation. To address the issue that using top-k feature vectors leads to a catastrophic decrease in utility, while using bottom feature vectors results in easy removal, we select the optimal point based on the following criteria:

$$\mathcal{I} = \{i \mid \tau_{\text{lower}} \cdot \lambda_1 \leq \lambda_i \leq \tau_{\text{upper}} \cdot \lambda_1\}. \tag{5}$$

Thresholds $\tau_{\text{lower}}, \tau_{\text{upper}}$ are determined via validation criteria ensuring functional integrity. The top-k indices from $\mathcal{I}$ form the functional backbone subspace $U^\star$.

4.4. Phase 4: Projection and Consistency

We introduce an anchored projection to define a stable coordinate system:

$$z(x) = U^{\star\top}(r(x;\theta) - \mu_{\mathcal{D}}) \in \mathbb{R}^k, \tag{6}$$

where $\mu_{\mathcal{D}}$ is the global mean representation from the frozen model f_{θ_0}. To preserve behavior, we impose a Vector Consistency constraint:

$$L_{\text{con}} = \mathbb{E}_{x \sim \mathcal{D}}\|z(x) - z_{\theta_0}(x)\|_2^2. \tag{7}$$

4.5. Phase 5: Watermark Embedding and Verification

The model owner generates a challenge set $\mathcal{C}$ and samples M mutually orthogonal secret key vectors $\{b_1, \ldots, b_M\}$ with $b_j \in \mathbb{R}^k$, where $b_i^\top b_j \approx 0$ for $i \neq j$. Binary messages are encoded into target signs $y_j \in \{-1, 1\}$.

Embedding Loss. We inject the watermark into the functional backbone using a margin-based objective:

$$L_{\text{wm}} = \mathbb{E}_{x \sim \mathcal{C}} \sum_{j=1}^{M} \max\left(0, \gamma - y_j \frac{b_j^\top z(x)}{\|b_j\|_2}\right), \tag{8}$$

where $\gamma > 0$ is the target margin. The watermarked model is obtained by minimizing the total objective:

$$L = L_{\text{LM}} + \lambda_{\text{wm}} L_{\text{wm}} + \lambda_{\text{con}} L_{\text{con}}. \tag{9}$$

Statistical Verification. Ownership is established by computing the aggregated detection score S:

$$S = \frac{1}{|\mathcal{C}|M} \sum_{x \in \mathcal{C}} \sum_{j=1}^{M} y_j \frac{b_j^\top z(x)}{\|b_j\|_2}. \tag{10}$$

Under the null hypothesis H_0, $S \sim \mathcal{N}(0, \sigma_0^2)$. The False Positive Rate (FPR) for a threshold T is bounded by:

$$\text{FPR} = \frac{1}{2}\text{erfc}\left(\frac{T}{\sqrt{2}\sigma_0}\right). \tag{11}$$

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

Table 1. Performance comparison of Functional Subspace Watermarking (FSW) across diverse LLM backbones. Our method achieves near-perfect bit accuracy and high detection scores while maintaining competitive functional utility.
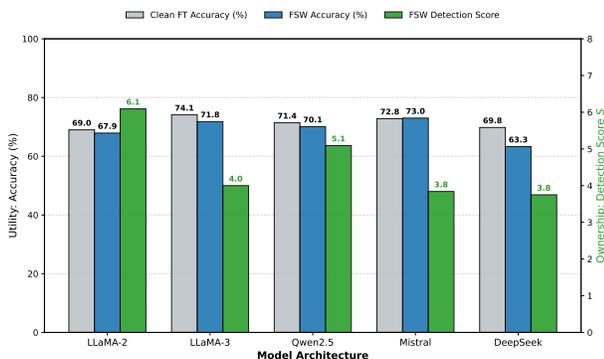

Figure 3. Functional Preservation and Ownership Detectability.

Multi-bit Decoding. For message recovery, per-bit statistics are computed as $S_j = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{b_j^\top z(x)}{\|b_j\|_2}$. The original bits are recovered via $\hat{y}_j = \text{sign}(S_j)$ and decoded through Error-Correcting Codes (ECC).

The complete procedure for watermark generation, embedding, and statistical verification is summarized in Algorithm 1 in the Appendix A.

5. Experiments

5.1. Experimental Setting

Models and Datasets. To evaluate the effectiveness and generality of FSW across diverse architectures, we select representative autoregressive language models including 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). We employ a multi-stage data strategy: WikiText-2 (Liu et al.) is utilized for estimating the Fisher matrix F and the compression invariance matrix $C_\mathcal{A}$; C4

(Zhu et al., 2023) is used for watermark embedding fine-tuning to improve distributional robustness. Downstream benchmarks including HellaSwag (Zellers et al., 2019) and ARC (Clark et al., 2018) are used to verify that watermark embedding does not degrade task-level capabilities. Detailed selection principles and dataset configurations are provided in Appendix B.

Baseline. To evaluate the superiority of FSW, we compared it with several state-of-the-art (SOTA) baseline methods and technical variants: (1) Clean FT ;(2) EmMark (Zhang & Koushanfar, 2024); (3) Weighted Quantization (Li et al., 2023); (4) Naive Top-k: a variant of our framework that selects the k directions with the highest eigenvalues λ_i from GEVP without applying an adaptive spectral truncation strategy.

Implementation Details. We extract the hidden representation of the last token from an intermediate layer ℓ as the representation vector. The invariance matrix $C_\mathcal{A}$ is estimated using a family of compression operators $\mathcal{A}$ simulating capacity bottlenecks (Linear Projection), precision loss (Quantization Noise), and feature sparsification (Structural Dropout). Subspace dimension k and spectral truncation thresholds are selected via validation set perplexity constraints to balance robustness and utility. Complete hardware and training configurations are detailed in Appendix B.3.

5.2. Main result

Functional Performance. Our experimental results demonstrate that FSW achieves exceptional functional preservation across diverse LLM architectures. As illustrated in Figure 3, the side-by-side comparison of downstream task accuracy between the Clean FT baseline (light grey bars) and FSW (blue bars) reveals minimal performance degradation.

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

Table 2. Robustness analysis of FSW under various model attacks. The results across multiple LLM backbones show that ownership signals remain recoverable even after significant parameter-level modifications.

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

Table 3. Impact of Error-Correcting Codes (ECC) on message decoding accuracy (Llama-2-7B-hf). The results demonstrate that `Hamming (7,4)` successfully eliminates residual errors at the 8-bit critical point, ensuring irrefutable ownership verification.

For instance, the PPL increment for Meta-LLaMA-3-8B is only 0.37, while maintaining a high average accuracy of 71.8%. Simultaneously, the green bars represent the high-magnitude detection scores S, showing a distinct separation from the null hypothesis across all backbones. This visualized method profile validates the efficacy of our spectral truncation strategy: by avoiding modifications to highly sensitive representation directions critical for task performance, FSW successfully embeds watermarks into a carrier space that is both robust and non-intrusive to the model's original capabilities.

Furthermore, the generalization capability of FSW across different training corpora is highlighted in Table 8 in Appendix C. Even when the embedding fine-tuning is shifted from the high-quality WikiText-2 dataset to the massive C4 corpus, the framework maintains consistent functional utility and robust detectability on the Meta-LLaMA-3-8B backbone. For instance, while using the C4 dataset slightly increases the perplexity (ΔPPL = +0.90), it yields a perfect bit accuracy of 100% and a near-ideal AUC of 1.000 . This suggests that the functional backbone identified by FSW captures fundamental task-relevant directions that are inherent to the model's architecture rather than being overfitted to a specific dataset, ensuring reliable ownership protection across diverse deployment environments .
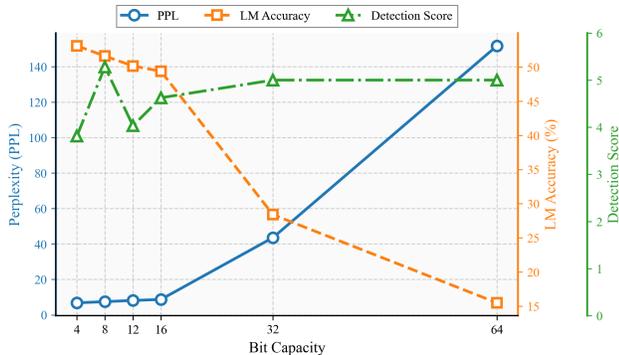
Figure 4. **Bit Capacity vs. Utility and Detectability**

Figure 5. **Robustness–Capacity Trade-off under Different Embedding Dimensions**

Statistical Verifiability. Experimental results validate the superior ownership verification capability and rigorous statistical reliability of FSW. Across all evaluated backbones, our method consistently yields detection scores S that significantly exceed the theoretical critical thresholds τ. As detailed in Table 7, we evaluate the model under various significance levels α. For DeepSeek-7B, the margin $S-\tau$ remains as high as 3.0587 even at an extremely stringent level of $\alpha = 10^{-8}$, effectively rejecting the null hypothesis H_0 with near-zero probability of false positives. While Mistral-7B shows a marginal failure at the extreme $\alpha = 10^{-8}$ level, it maintains a robust positive margin of 0.3776 at $\alpha = 10^{-6}$, confirming high detection confidence. This demonstrates that the functional subspace constructed via GEVP identifies carrier directions with an exceptionally high signal-to-noise ratio (SNR), enabling irrefutable ownership proof that is statistically distinguishable from random fluctuations.

5.3. Robustness Analysis

Robustness to Parameter Perturbations. The resilience of FSW against post-hoc modifications is demonstrated by its performance under parameter-level attacks such as LoRA fine-tuning, quantization, and pruning. As shown in Table 2, the framework maintains, and in some cases even enhances, its detection capabilities under these perturbations; for instance, the detection score for LLaMA-2-7B-hf increased from a baseline of 6.09 to 6.56 after INT4 quantization and 6.72 after magnitude pruning. Furthermore, the watermark proved resistant to LoRA fine-tuning on Meta-LLaMA-3-8B, maintaining an 87.5% bit accuracy while slightly increasing the detection score from 4.00 to 4.22. This stability, further supported by the distinct separation of detection scores in Figure 3, confirms that FSW successfully anchors ownership signals into a functional backbone that remains invariant to common precision losses and structural modifications.

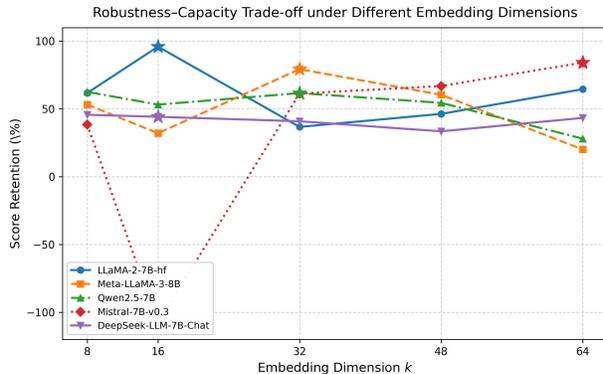**Resistance to Model Distillation.** FSW exhibits superior robustness against model distillation attacks compared to current state-of-the-art methods by explicitly decoupling task-critical sensitivity from compression-induced variance. According to Table 5, while baseline methods like EmMark often suffer from significant utility degradation, FSW preserves the original model performance, evidenced by the stable PPL of 5.91 for LLaMA-2-7B after distillation. The "Backbone Distillation" results in Table 2 further highlight this efficacy, where DeepSeek-LLM-7B-Chat achieved a perfect 100% bit accuracy even after its weights were reparameterized. This statistical reliability is underscored in Table 5 in Appendix C, where the detection margin for DeepSeek-7B remained as high as 3.0587 at an extreme significance level of $\alpha = 10^{-8}$, providing irrefutable proof of ownership even under high-compression scenarios.

5.4. Ablation Studies.

Bit Capacity To determine the optimal payload for ownership verification, we analyze the trade-off between watermark capacity, model utility, and detection reliability. As illustrated in Figure 5, FSW maintains a stable equilibrium when the bit capacity $B \leq 16$, where the PPL remains nearly constant and the LM Accuracy stays at a high level (above 45%). However, increasing the capacity to 32 or 64 bits leads to a limitation, where the PPL surges to over 150 and accuracy drops sharply, indicating a structural saturation of the functional subspace. This capacity threshold is further validated by the generalization results in Table 3, where a configuration within this robust range (typically 16 bits) achieves a perfect Bit Accuracy of 100% and an AUC of 1.000 on the C4 dataset, despite the increased complexity of the corpus. Notably, while excessive embedding distorts model weights, the Detection Score remains consistently high across all evaluated capacities, confirming that the underlying signal remains recoverable even under high-load conditions. Consequently, 16 bits is identified as the optimal capacity, as it maximizes the information payload for ECC without compromising the model's original functional

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 Thres.	4.91	-2.73	-7.64	87.5%	12.5%	-75.0%	7.16	6.85	+0.30

Table 4. Ablation study of FSW components on Llama-2-7B-hf. *Pre* and *post* denote metrics before and after the robustness attack. Each variant systematically removes a key design element (Compression Invariance, Anchored Projection, Consistency Loss, or Adaptive Thresholding) to evaluate its individual contribution to robustness and utility.

integrity.

Ablation Study of Components. The ablation study conducted on the Llama-2-7B-hf model demonstrates that the integration of all FSW components is essential for maintaining the optimal balance between watermark robustness and model utility. As detailed in Table 4, the Full FSW configuration consistently outperforms all variants, maintaining a stable PPL of 5.91 (ΔPPL = 0.00) and a robust post-attack bit accuracy of 87.5%. In contrast, the removal of the Adaptive Thresholding (w/o Adaptive Thres.) results in the most severe performance degradation, with the post-attack bit accuracy plummeting to 12.5% and the detection score S dropping to a negative value of -2.73. These results indicate that while Compression Invariance and Consistency Loss are vital for signal persistence, the Anchored Projection and Adaptive Thresholding serve as the foundational pillars for ensuring a stable recovery coordinate system and protecting task-critical representation directions from watermark-induced distortion.

Ablation Study of Embedding Dimension k. The ablation analysis of the embedding dimension k reveals that selecting an optimal subspace size is crucial for maximizing watermark retention without sacrificing model performance. As illustrated in Table 9 in Appendix C, Table 6 (in Appendix C), and Figure 5, the optimal k varies significantly across architectures to balance watermark capacity and structural stability, with Llama-2-7B reaching its peak score retention of 95.90% at $k = 16$, while larger models like Mistral-7B require a higher dimension of $k = 64$ to achieve a maximum retention rate of 84.15%. Data from the comprehensive scan in Table 9 indicates that deviations from these optimal values lead to a significant performance trade-off: smaller dimensions often fail to capture sufficient carrier directions, resulting in lower detection scores after robustness attacks, whereas excessively large dimensions begin to overlap with highly sensitive representation regions, causing an unnecessary increase in perplexity. These findings suggest that the optimal k serves as a structural bottleneck that effectively encapsulates the stable functional backbone of each specific LLM, ensuring that the embedded ownership sig-

nals are both robust to parameter-level perturbations and non-intrusive to the model's original utility.

5.5. Cost Analysis

Training Overhead Analysis. The evaluation of training overhead reveals that while FSW introduces additional computational requirements compared to standard fine-tuning, the overall costs remain within a manageable range for 7B-scale models. As detailed in Table 10, the total training time for the FSW pipeline on Qwen2.5-7B is 1855.68 seconds across 8001 steps, representing a reasonable increase over the 1101.55 seconds required for standard fine-tuning. Most of this overhead is concentrated in the Phase 4: Watermark Embedding stage, which takes 1764.38 seconds and requires significantly higher GPU memory (56410.34 MB) compared to the baseline (29820.63 MB). This increase in resource consumption is primarily attributed to the dual-model consistency constraints employed during the embedding phase, which necessitate maintaining additional representation states to ensure that the watermarking process does not degrade the original model's functional utility.

6. Conclusion

This paper presents the `FSW`, a framework for LLMs ownership protection. To address watermark failure caused by fine-tuning, quantization, and distillation, we construct a low-dimensional functional subspace that combines task-criticality with compression-invariance by solving the GEVP. This method embeds the watermark into the model's core representation directions to ensure robustness under parameter-level modifications. To balance robustness and model utility, we introduce an adaptive spectral truncation strategy to avoid performance degradation and utilize vector consistency constraints to preserve the original semantic distribution. Extensive experiments have shown that `FSW` effectively defends against various model attacks while maintaining an extremely low false positive rate. This provides a rigorous technical approach for the intellectual property protection of large-scale models.

Impact Statement

This work aims to advance research on model-side watermarking for large language models by improving robustness under common model modifications. The proposed method is intended to support model ownership verification and provenance analysis. We do not anticipate significant negative societal impacts arising from this work.

References

Adi, Y., Baum, C., Cisse, M., Pinkas, B., and Keshet, J. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pp. 1615–1631, 2018.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Dabiriaghdam, A. and Wang, L. SimMark: A robust sentence-level similarity-based watermarking algorithm for large language models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (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/.

Dai, Y., Li, Z., Ji, Z., and Wang, S. Seal: Subspace-anchored watermarks for llm ownership. *arXiv preprint arXiv:2511.11356*, 2025.

Darvish Rouhani, B., Chen, H., and Koushanfar, F. Deep-signs: 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.

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035): 818–823, 2024.

He, H., Liu, Y., Wang, Z., Mao, Y., and Bu, Y. Theoretically grounded framework for llm watermarking: A distribution-adaptive approach. In *The 1st Workshop on GenAI Watermarking*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Li, L., Jiang, B., Wang, P., Ren, K., Yan, H., and Qiu, X. Watermarking llms with weight quantization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3368–3378, 2023.

Li, Z., Wu, D., Wang, S., and Su, Z. 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.

Liu, A. Z., Paquette, E., and Sous, J. Evolution of the spectral dimension of transformer activations. In *OPT 2025: Optimization for Machine Learning*.

Liu, Y., Zhao, X., Kruegel, C., Song, D., and Bu, Y. In-context watermarks for large language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Mao, M., Wei, D., Chen, Z., Fang, X., and Chau, M. 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.

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.

Niess, G. and Kern, R. 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.

Pasquini, D., Kornaropoulos, E. M., and Ateniese, G. {LLMmap}: Fingerprinting for large language models.

In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 299–318, 2025.

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

Sander, T., Fernandez, P., Durmus, A., Douze, M., and Furon, T. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37:21079–21113, 2024.

Shao, S., Li, Y., He, Y., Yao, H., Yang, W., Tao, D., and Qin, Z. Sok: Large language model copyright auditing via fingerprinting. *arXiv preprint arXiv:2508.19843*, 2025.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

Xu, J., Wang, F., Ma, M., Koh, P. W., Xiao, C., and Chen, M. 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.

Xu, X., Jia, J., Yao, Y., Liu, Y., and Li, H. Robust multi-bit text watermark with llm-based paraphrasers. In *Forty-second International Conference on Machine Learning*, 2025.

Yamabe, S., Waseda, F. K., Takahashi, T., and Wataoka, K. 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.

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

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan,

Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yao, H., Lou, J., Qin, Z., and Ren, K. Promptcare: Prompt copyright protection by watermark injection and verification. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 845–861. IEEE, 2024.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Zhang, R. and Koushanfar, F. 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.

Zhang, R., Hussain, S. S., Neekhara, P., and Koushanfar, F. {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.

Zhu, W., Hessel, J., Awadalla, A., Gadre, S. Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W. Y., and Choi, Y. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

A. Algorithm.

Algorithm 1 Functional Subspace Watermarking (FSW) Implementation Pipeline

---

1: **Input:** Pre-trained model f_{θ_0}, calibration set $\mathcal{D}$, challenge set $\mathcal{C}$, layer ℓ, subspace dimension k, margin γ, target bits $\{y_j\}_{j=1}^{M}$.

2: **Phase 1: Representation Analysis**

3:     Extract hidden states $r(x; \theta_0)$ from layer ℓ for $x \sim \mathcal{D}$ and compute mean $\mu_{\mathcal{D}}$.

4:     Estimate Fisher matrix F (Eq. 2) and Invariance matrix $C_{\mathcal{A}}$ (Eq. 3) via stochastic operators $\mathcal{A}$.

5: **Phase 2: Subspace Construction & Optimization**

6:     Solve Generalized Eigenvalue Problem (GEVP): $Fu_i = \lambda_i C_{\mathcal{A}} u_i$ to obtain eigenvectors $\{u_i\}$.

7:     Determine spectral sweet spot $\mathcal{I}$ via adaptive truncation thresholds $\tau_{\text{lower}}, \tau_{\text{upper}}$ (Eq. 5).

8:     Construct the functional backbone projection matrix $U^{\star} = [u_{i_1}, \ldots, u_{i_k}]$ where $i \in \mathcal{I}$.

9: **Phase 3: Constrained Watermark Embedding**

10:     Generate M mutually orthogonal secret key vectors $\{b_1, \ldots, b_M\} \in \mathbb{R}^k$.

11:     Fine-tune model f_{θ} by minimizing the joint objective (Eq. 10):

12:        $L = L_{\text{LM}} + \lambda_{\text{wm}} L_{\text{wm}} + \lambda_{\text{con}} L_{\text{con}}$.

13: **Phase 4: Statistical Ownership Verification**

14:     Given a suspected model, project representations into $U^{\star}$ and compute S (Eq. 11).

15:     Perform hypothesis test: reject H_0 if FPR $= \frac{1}{2}\text{erfc}\left(\frac{T}{\sqrt{2}\sigma_0}\right) < \alpha$ (threshold α).

16: **Phase 5: Multi-bit Message Decoding**

17:     Compute per-bit projection statistics $S_j = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{b_j^{\top} z(x)}{\|b_j\|_2}$ for $j = 1, \ldots, M$.

18:     **Return:** Extracted bits $\hat{y}_j = \text{sign}(S_j)$ and final ECC-decoded message.

B. Detailed Experimental Setup

B.1. Model Selection and Specifications

The model selection is shown in Table 11 in Appendix C.

B.2. Dataset Descriptions and Allocation

We combine multiple datasets for subspace estimation, fine-tuning, and functional evaluation to ensure that this is not a coincidence.

- **WikiText-2:** A high-quality collection of Wikipedia articles. We use 500 samples for estimating the Fisher matrix F and invariance matrix $C_{\mathcal{A}}$, and 500 separate samples for evaluating language modeling perplexity (PPL).

- **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.

B.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 11. For the invariance matrix $C_{\mathcal{A}}$, we apply $N = 3$ stochastic perturbations per sample using a family of operators $\mathcal{A}$:

- **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 $p = 0.9$.

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%	—	—	—
	Full DIFSW	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%	—	—	—
	Full DIFSW	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%	—	—	—
	Full DIFSW	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%	—	—	—
	Full DIFSW	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%	—	—	—
	Full DIFSW	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 5. **Baseline Comparison under Model Distillation.** We report perplexity (PPL) and downstream accuracies (HellaSwag and ARC-Challenge) before and after distillation for multiple watermarking baselines.

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

C. Supplementary Table

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 6. 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	Significance Level (α)	Threshold T_α	Detection Score S	Margin ($S - T_\alpha$)	Detected
Mistral-7B	1e−02	1.5838	3.8438	2.2600	✓
	1e−03	2.1762	3.8438	1.6675	✓
	1e−04	2.6639	3.8438	1.1798	✓
	1e−06	3.4662	3.8438	0.3776	✓
	1e−08	4.1321	3.8438	-0.2883	✗
DeepSeek-7B	1e−02	0.2358	3.7500	3.5142	✓
	1e−03	0.3417	3.7500	3.4083	✓
	1e−04	0.4288	3.7500	3.3212	✓
	1e−06	0.5723	3.7500	3.1777	✓
	1e−08	0.6913	3.7500	3.0587	✓

Table 7. Detection Thresholds and Significance Test Results across Different Significance Levels α. FSW successfully rejects the null hypothesis H_0 with a substantial positive margin even under extreme significance constraints (e.g., $\alpha = 10^{-8}$ for DeepSeek).

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

Table 8. Generalization across different training datasets on Meta-LLaMA-3-8B. The results demonstrate that FSW maintains consistent functional utility and high detectability regardless of the fine-tuning corpus used.

Model	Candidates	Opt. k	Pre-Attack (Original)			Post-Attack (Robustness)			Ret. (%)
			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%	**95.90**
LLaMA-3-8B	$\{8, 16, 32, \dots\}$	32	6.20	4.00	100.0%	6.12	3.17	62.5%	**79.30**
Qwen2.5-7B	$\{8, 16, 32, \dots\}$	32	7.24	5.09	100.0%	7.66	3.14	75.0%	**61.66**
Mistral-7B-v0.3	$\{8, 16, 32, \dots\}$	64	17.25	3.84	100.0%	6.59	3.23	75.0%	**84.15**
DeepSeek-7B-Chat	$\{8, 16, 32, \dots\}$	16	77.31	3.75	100.0%	61.39	1.66	100.0%	**44.17**

Table 9. Optimal embedding dimension k selection across backbone models. The optimal k is determined by maximizing the retention rate (Ret.%) of detection scores after robustness attacks, ensuring an effective trade-off between watermark capacity and structural stability.

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**

Table 10. 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 within a manageable range for 7B-scale models.

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

Table 11. 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.