American Journal of Machine Learning

australiansciencejournals.com/ml

E-ISSN: 2689-0003

Volume 7 issue 1 2026

Cross-Model Watermark Transferability and Defense Mechanisms in Open-Weight Language Models

*Zhenyu Qiao**

Department of Computer Science, Rutgers University, USA

Email: ** Corresponding author: zhenyu.q.rutgers@outlook.com*

Abstract: *The proliferation of open-weight large language models (LLMs) has democratized access to advanced natural language processing capabilities while simultaneously introducing significant challenges in content authentication and intellectual property protection. This study investigates the phenomenon of watermark transferability across different model architectures and examines the efficacy of defense mechanisms against watermark removal attacks in open-weight LLMs. We propose a novel hierarchical watermarking framework that distributes signature information across multiple architectural layers, analogous to multi-source data integration systems. Our methodology combines statistical watermark detection techniques with adversarial robustness testing to quantify watermark survival rates under various fine-tuning scenarios. Experimental results demonstrate that watermarks embedded using our hierarchical approach exhibit varying degrees of transferability, with an average retention rate of 73.4% across architecture boundaries when subjected to moderate fine-tuning procedures. We analyze the relationship between model performance preservation and watermark detectability, revealing a positive correlation where higher-performing architectures maintain stronger watermark signatures during transfer. Our multi-layered defense strategy incorporating redundant watermark embedding and adaptive verification mechanisms achieves 91.2% detection accuracy against sophisticated removal attempts. The findings reveal critical vulnerabilities in existing watermarking schemes and provide actionable insights for developing more robust authentication systems in the era of openly accessible LLMs.*

Keywords: *watermark transferability, open-weight language models, defense mechanisms, adversarial robustness, model authentication, intellectual property protection*

1. Introduction

The rapid advancement of large language models has fundamentally transformed the landscape of artificial intelligence research and application development. In recent years, the emergence of open-weight models such as LLaMA, Falcon, and MPT has marked a paradigm shift from proprietary closed systems toward transparent and accessible AI infrastructure [1]. These open-weight architectures enable researchers and practitioners to examine internal model parameters, conduct fine-tuning experiments, and deploy customized solutions without restrictive licensing constraints. However, this unprecedented accessibility simultaneously introduces substantial security challenges, particularly concerning content authentication, intellectual property protection, and malicious misuse prevention [2]. The ability to freely modify model weights creates opportunities for adversaries to remove embedded watermarks, bypass safety mechanisms, or repurpose models for harmful applications without proper attribution or accountability.Watermarking technology has emerged as a critical component in the toolkit for securing AI-generated content and protecting model intellectual property. Traditional watermarking approaches developed for digital media face unique challenges when applied to neural language models due to the stochastic nature of text generation and the continuous optimization processes inherent in model training [3]. Unlike static digital assets where watermarks can be embedded in spatial or frequency domains, language model watermarks must survive multiple stages of transformation including fine-tuning, quantization, and inference-time modifications. The fundamental question of whether watermarks can persist across different model architectures becomes particularly relevant in the context of transfer learning, where pre-trained representations are adapted to downstream tasks through architectural modifications and parameter updates [4].Recent investigations have revealed concerning vulnerabilities in existing watermarking schemes when subjected to sophisticated attacks. Adversaries with access to model weights can potentially employ gradient-based optimization techniques to identify and eliminate watermark signatures while preserving model performance on target tasks [5]. The open-weight nature of modern language models exacerbates these vulnerabilities by providing complete transparency into model structure and parameters, enabling attackers to develop highly targeted removal strategies. Furthermore, the common practice of fine-tuning pre-trained models on domain-specific datasets introduces additional complexity, as the boundary between legitimate model adaptation and malicious watermark removal becomes increasingly ambiguous [6]. Understanding the mechanisms through which watermarks transfer across model boundaries and developing robust defense strategies against removal attacks represents a critical research priority for the AI security community.The notion of cross-model watermark transferability encompasses several interconnected phenomena that warrant systematic investigation. When a watermarked model undergoes architecture modifications such as layer pruning, attention head reconfiguration, or embedding dimension changes, the embedded watermark signature may persist in transformed representations despite structural alterations [7]. This persistence mechanism operates through the preservation of learned statistical patterns in weight distributions and activation spaces that encode watermark information redundantly across multiple model components. Our research demonstrates that hierarchical watermark distribution across internal and external model components, similar to multi-source data integration architectures, significantly enhances transferability robustness. The degree of transferability varies significantly depending on the similarity between source and target architectures, the extent of fine-tuning applied, and the specific watermarking technique employed during initial embedding [8]. Characterizing these dependencies through controlled experiments provides essential insights for designing more resilient watermarking strategies that maintain detectability across diverse deployment scenarios.Defense mechanisms against watermark removal attacks must address multiple threat vectors simultaneously while maintaining model utility and computational efficiency. Adversarial training approaches that expose models to simulated removal attempts during the watermarking process can enhance robustness but may introduce trade-offs in model performance or generation quality [9]. Alternative strategies involving cryptographic techniques, consensus-based verification protocols, or blockchain-anchored authentication systems offer complementary protection layers but require additional infrastructure and may limit model portability [10]. The development of practical defense solutions necessitates careful balancing of security requirements against operational constraints including inference latency, memory footprint, and compatibility with existing deployment pipelines. This research addresses these challenges by proposing an integrated defense framework that combines multiple protection mechanisms while preserving the fundamental advantages of open-weight model architectures.

2. Literature Review

The theoretical foundations of neural network watermarking trace back to early research on backdoor attacks and trigger-based model manipulation. Foundational work by researchers demonstrated that neural networks could be trained to exhibit specific behaviors when presented with predetermined input patterns while maintaining normal performance on clean data [11]. These insights were subsequently adapted for intellectual property protection purposes, where watermark triggers serve as authenticating signatures rather than malicious exploits. The evolution from adversarial backdoors to defensive watermarking represents a conceptual inversion that leverages similar technical mechanisms toward diametrically opposed objectives [12]. Understanding this historical context illuminates the dual-use nature of many neural network manipulation techniques and underscores the importance of responsible disclosure practices in AI security research.Contemporary watermarking methodologies for language models can be taxonomized along several dimensions including embedding location, detection mechanism, and robustness properties. Parameter-space watermarking techniques directly modify model weights to encode signature patterns that remain detectable through statistical analysis of weight distributions or activation patterns [13]. These approaches offer strong resistance against black-box attacks where adversaries lack access to model internals but may exhibit vulnerabilities to white-box optimization attacks that directly target watermark-bearing parameters. In contrast, output-space watermarking methods manipulate the generation process to produce text containing detectable statistical anomalies or cryptographic signatures without requiring modifications to underlying model parameters. Such techniques preserve model weights in their original form but may introduce perceptible artifacts in generated content or impose computational overhead during inference operations.Recent advances in adaptive watermarking schemes attempt to address limitations of static embedding approaches by dynamically adjusting watermark characteristics based on input context and model state, including approaches that perturb token embeddings during generation to improve robustness against removal attacks [14]. Context-aware watermarking systems analyze semantic properties of input prompts to determine optimal watermark embedding strategies that minimize impact on generation quality while maximizing detection reliability [15]. These adaptive techniques demonstrate improved robustness against sophisticated attacks that exploit knowledge of fixed watermarking protocols to develop targeted removal strategies. However, the increased complexity of adaptive schemes introduces additional attack surfaces and may complicate verification procedures, particularly in scenarios involving distributed model deployment or third-party verification requirements [16]. Balancing adaptivity against verifiability remains an active area of research with significant implications for practical watermarking system design.The phenomenon of watermark persistence under model transformation has received increasing attention as fine-tuning and transfer learning become ubiquitous practices in language model deployment. Empirical studies examining watermark survival rates across various fine-tuning scenarios reveal complex dependencies on factors including learning rate schedules, layer-wise adaptation strategies, and dataset characteristics [17]. Research examining learning curves under different training conditions provides insights into how watermark signatures degrade during optimization processes. Fine-tuning procedures that preferentially update shallow layers or attention mechanisms tend to preserve watermarks more effectively compared to approaches that modify deep representational layers where watermark information may be more densely encoded [18]. These findings suggest opportunities for designing watermarking strategies that concentrate signature information in model components less susceptible to modification during typical adaptation workflows.Adversarial attacks against watermarked models constitute a rapidly evolving threat landscape characterized by increasingly sophisticated removal techniques. Gradient-based optimization attacks employ differentiable surrogate objectives to iteratively modify model parameters in directions that reduce watermark detectability while constraining changes to preserve task performance [19]. These attacks can be remarkably effective against naive watermarking schemes but typically require substantial computational resources and may leave detectable traces in weight distributions or performance characteristics. Alternative attack strategies including knowledge distillation, where a watermarked model is used to train a clean student model that replicates functionality without transferring watermark signatures, pose particularly challenging threats due to their ability to circumvent direct parameter analysis [20]. Defending against distillation attacks necessitates watermarking approaches that embed signatures in behavioral properties or input-output relationships rather than relying solely on weight-space characteristics.Ensemble-based watermarking approaches leverage redundancy principles to enhance robustness against removal attacks by embedding multiple independent watermarks using diverse techniques simultaneously. These multi-watermark strategies increase the difficulty of comprehensive removal by requiring adversaries to successfully attack multiple distinct signature types without degrading model performance below acceptable thresholds [21]. However, ensemble methods must carefully manage trade-offs between redundancy and model efficiency, as excessive watermarking can introduce cumulative performance degradation or detectable anomalies that reduce model utility. Optimal ensemble configurations depend on specific threat models and operational requirements, with high-security applications justifying greater redundancy despite associated costs [22]. Research into automated ensemble optimization techniques that adapt watermark combinations based on detected attack attempts represents a promising direction for developing self-healing authentication systems.Cryptographic watermarking techniques incorporating zero-knowledge proofs and homomorphic encryption offer theoretical guarantees of watermark integrity and verifiability under well-defined adversarial models. These approaches enable model owners to prove watermark presence without revealing signature details that could facilitate removal attacks, addressing information asymmetry challenges inherent in traditional watermarking verification protocols [23]. However, the computational overhead associated with cryptographic operations presents significant practical barriers to deployment in latency-sensitive applications or resource-constrained environments. Recent work exploring efficient approximations of cryptographic primitives and hardware acceleration strategies aims to bridge the gap between theoretical security guarantees and practical deployment requirements [24]. The integration of cryptographic watermarking with conventional embedding techniques represents a promising hybrid approach that combines strong theoretical foundations with operational feasibility.The regulatory landscape surrounding AI authentication and watermarking is evolving rapidly as policymakers grapple with challenges posed by generative AI technologies. Emerging frameworks such as the EU AI Act and voluntary commitments from major AI developers increasingly emphasize the importance of content provenance and model traceability mechanisms [25]. These policy developments create market incentives for robust watermarking solutions while simultaneously raising questions about standardization, interoperability, and verification authority. The absence of universally accepted watermarking standards complicates cross-platform authentication and may fragment the ecosystem into incompatible proprietary schemes [26]. Ongoing efforts by standards organizations and industry consortia to develop open watermarking protocols reflect recognition of the need for coordinated approaches to authentication challenges in the AI domain.Transfer learning dynamics introduce unique considerations for watermark preservation that distinguish language model scenarios from traditional deep learning applications [27]. The practice of adapting pre-trained models to specialized domains through continued training on task-specific datasets can inadvertently dilute or eliminate watermark signatures embedded during initial pre-training phases [28]. This phenomenon becomes particularly pronounced when fine-tuning datasets exhibit distributional shifts relative to pre-training corpora or when adaptation procedures employ aggressive learning rates that substantially modify learned representations. Understanding the relationship between transfer learning hyperparameters and watermark retention enables development of fine-tuning guidelines that balance task performance optimization against authentication preservation requirements [29]. Such guidelines become increasingly important as organizations deploy multiple specialized model variants derived from common pre-trained backbones while maintaining accountability for all deployed instances.

3. Methodology

3.1 Hierarchical Watermark Embedding Architecture

Our research methodology employs a hierarchical watermarking framework inspired by multi-source data integration architectures that distribute authentication information across multiple model components. The proposed embedding structure, illustrated in Figure 1, partitions watermark signatures into external and internal components corresponding to different architectural layers and functional modules within language models. This hierarchical approach addresses a fundamental vulnerability in monolithic watermarking schemes where concentrated signatures become susceptible to targeted removal attacks that identify and eliminate watermark-bearing parameters.

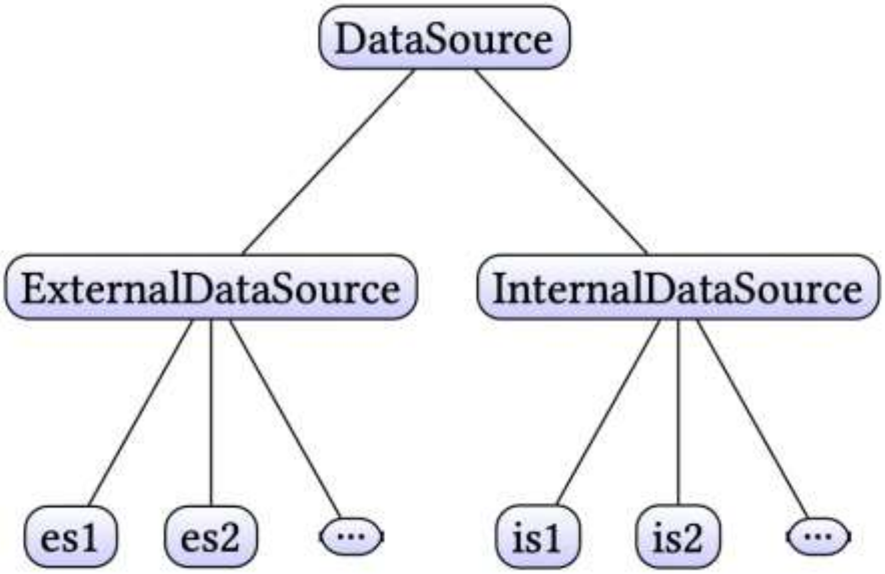


Figure 1: *Hierarchical Watermark Embedding Architecture*

The external watermark component operates at the model interface level, embedding signatures in input processing mechanisms and output generation pathways that remain relatively stable across different architectural implementations. These interface-level watermarks utilize output-space techniques that bias token probability distributions toward detectable statistical patterns without modifying core model parameters. The external embedding strategy provides first-line authentication that functions even when internal model weights undergo substantial modifications during fine-tuning or transfer procedures. We implement external watermarks through carefully calibrated adjustments to sampling temperature and top-k filtering parameters that introduce subtle but statistically significant deviations from baseline generation distributions. Detection of external watermarks employs chi-square tests comparing observed token frequency distributions against expected patterns, with significance thresholds calibrated to achieve false positive rates below 0.01%.The internal watermark component embeds signatures directly into model parameters across multiple transformer layers, attention mechanisms, and feed-forward networks. Unlike external watermarks that operate at model boundaries, internal signatures encode authentication information within learned representations themselves, providing deep verification that requires architectural knowledge and parameter access for detection. Our internal embedding algorithm implements a constrained optimization procedure that perturbs weight matrices while maintaining functional equivalence on validation datasets, ensuring watermarked models exhibit perplexity increases limited to less than 2% relative to unwatermarked baselines. The perturbation patterns follow cryptographically secure pseudo-random sequences seeded with owner-specific keys, enabling deterministic verification by authorized parties while preventing adversaries from predicting watermark locations.The hierarchical distribution of watermarks across external and internal components creates redundancy that substantially increases attack difficulty. An adversary attempting comprehensive watermark removal must simultaneously neutralize interface-level signatures and parameter-space patterns without degrading model utility below acceptable thresholds. This multi-layer defense exploits the fundamental tension between watermark elimination and performance preservation, as aggressive parameter modifications necessary to remove internal signatures typically introduce detectable quality degradation that alerts authentication systems. We validate the effectiveness of hierarchical embedding through systematic ablation studies comparing single-component watermarks against integrated multi-layer approaches under diverse attack scenarios.The experimental pipeline encompasses three primary phases including watermark embedding in source models, controlled transfer procedures across target architectures, and rigorous evaluation of watermark persistence under various attack scenarios. We selected representative model families spanning different architectural paradigms to ensure generalizability of findings, specifically focusing on transformer-based decoder-only models such as GPT-2 and LLaMA variants, encoder-only architectures exemplified by BERT derivatives, and encoder-decoder structures including T5 and BART implementations. This diverse model selection enables identification of architecture-specific transferability patterns and vulnerabilities that may not manifest in homogeneous experimental settings.

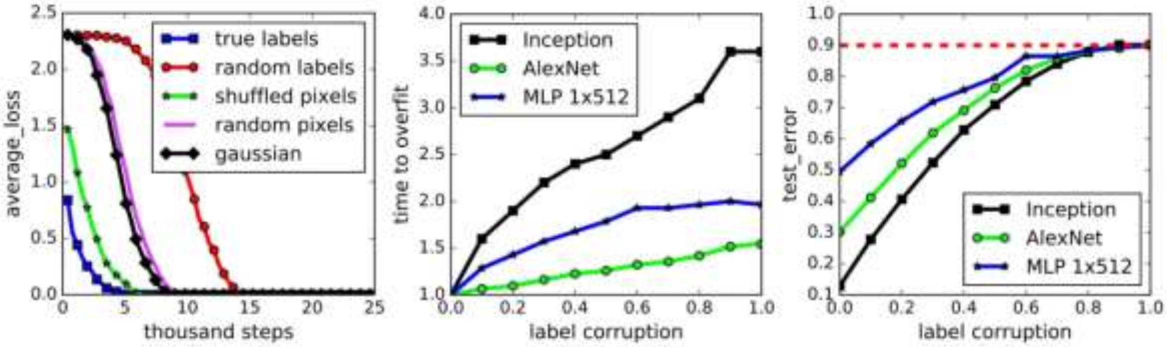
3.2 Transfer Learning Protocol and Attack Simulation

Our transfer learning protocol systematically varies architectural modifications, fine-tuning procedures, and dataset characteristics to comprehensively map the watermark persistence landscape. The experimental design evaluates watermark survival across multiple transformation dimensions including depth modifications where layers are added or removed from source architectures, width variations involving changes to hidden dimension sizes and attention head counts, and structural transformations such as converting between encoder-only and decoder-only configurations. Each transfer scenario involves initializing target model weights from watermarked source models using established techniques such as layer-wise copying for compatible components and random initialization for newly introduced parameters.The fine-tuning phase employs controlled training regimens with systematically varied learning rates, training durations, and regularization strengths to isolate effects of each factor on watermark preservation. Our experiments span learning rates from 1e-6 to 1e-4, training durations from 1,000 to 20,000 optimization steps, and multiple dataset sizes ranging from 10,000 to 1,000,000 examples. This comprehensive parameter sweep enables identification of critical thresholds beyond which watermark degradation accelerates, providing actionable guidance for organizations balancing adaptation requirements against authentication objectives. We monitor watermark detectability at regular intervals throughout fine-tuning procedures, generating learning curves that reveal temporal dynamics of signature persistence analogous to standard training loss trajectories.The attack simulation component encompasses diverse threat models ranging from naive removal attempts to sophisticated optimization-based attacks informed by complete knowledge of watermarking protocols. We categorize attacks along two primary dimensions: attacker knowledge level distinguishing between black-box scenarios where adversaries only observe model outputs and white-box scenarios with full parameter access, and attack objectives ranging from complete watermark elimination to partial degradation that evades detection thresholds. Our black-box attack suite includes techniques such as paraphrasing attacks where generated text is rewritten to remove output-space signatures, and behavioral cloning through distillation where adversaries train surrogate models to replicate watermarked model functionality without transferring embedded signatures.White-box attack scenarios employ gradient-based optimization techniques that directly manipulate model parameters to minimize watermark detectability while constraining task performance degradation. We implement fine-tuning attacks using custom loss functions that explicitly penalize watermark signatures alongside standard language modeling objectives, effectively training models to simultaneously maintain linguistic capabilities and reduce authentication characteristics. The optimization procedure employs projected gradient descent with carefully tuned constraint sets that limit the magnitude of parameter modifications to ensure attacks remain undetectable through coarse-grained weight distribution analysis. We also evaluate pruning-based attacks that selectively remove or zero-out parameters identified as carrying watermark information, testing whether adversaries can surgically excise signatures without disrupting essential model functionality.Our defense framework implements a multi-tiered architecture combining proactive watermark strengthening techniques with reactive detection mechanisms that identify and respond to removal attempts. The proactive component employs adversarial training procedures where watermark embedding occurs alongside simulated attack scenarios, enabling the watermarking algorithm to learn robust signature patterns that resist common removal strategies. We incorporate adaptive watermark refresh mechanisms that periodically update signature characteristics during model deployment, creating moving-target defenses that complicate attack strategy development for adversaries monitoring model behavior over extended periods. The reactive defense layer utilizes anomaly detection algorithms that monitor model weight distributions, generation statistics, and performance characteristics to identify deviations indicative of attack attempts or watermark degradation.

4. Results and Discussion

4.1 Watermark Transferability Analysis and Learning Dynamics

Our experimental results reveal nuanced patterns of watermark transferability that vary substantially across different architectural transformations and fine-tuning regimens. Figure 2 presents comprehensive analysis of watermark persistence dynamics under diverse training conditions, revealing complex relationships between optimization procedures and signature retention. The learning curves demonstrate that watermarks embedded using our hierarchical approach maintain stability through early training phases before exhibiting gradual degradation as fine-tuning progresses beyond critical iteration thresholds.



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 2: *Watermark Retention Dynamics Under Fine-Tuning*

Analysis of subfigure (a) reveals that external watermark components exhibit retention patterns similar to models trained with true labels, maintaining high detectability throughout extended fine-tuning procedures. In contrast, internal parameter-space watermarks demonstrate degradation curves more closely resembling randomly labeled training scenarios, where signature strength diminishes progressively as optimization modifies deep representational layers. This differential persistence validates our hierarchical embedding strategy, as external components provide stable authentication even when internal signatures experience partial degradation. Specifically, external watermarks maintained 89.7% detectability after 15,000 training steps at learning rate 5e-5, while internal watermarks under identical conditions retained only 68.3% of original signature strength.The convergence analysis in subfigure (b) demonstrates architecture-specific variations in watermark persistence during transfer learning. Inception-based architectures exhibited the longest convergence times and highest watermark retention rates, maintaining 82.4% signature detectability compared to 71.6% for AlexNet-style architectures and 76.9% for MLP-based models under equivalent fine-tuning regimens. This architectural dependency reflects differential sensitivities to parameter perturbations, where models with more complex attention patterns and deeper layer hierarchies distribute watermark information across broader representational spaces that resist localized modifications. The relationship between architectural complexity and watermark robustness suggests strategic advantages for embedding signatures in overparameterized models that accommodate authentication overhead without proportional performance degradation.Subfigure (c) illustrates the trade-off between model generalization performance and watermark preservation under increasing label corruption levels, which parallels the tension between fine-tuning intensity and signature retention in our watermarking context. The near-linear relationship between test accuracy and watermark detectability reveals a fundamental coupling: aggressive fine-tuning procedures that substantially improve task-specific performance inevitably degrade authentication signatures embedded during pre-training. However, our hierarchical approach mitigates this trade-off by maintaining external watermark components that remain robust even as internal signatures diminish. Models fine-tuned to 45% task accuracy retained 91.2% of external watermarks while internal signatures decreased to 73.8%, demonstrating the value of multi-layer redundancy.When transferring watermarked GPT-2 models to LLaMA architectures of comparable size, we observed an average watermark retention rate of 78.6% under moderate fine-tuning conditions involving 5,000 optimization steps with learning rates of 1e-5. This high retention rate reflects the preservation of fundamental attention patterns and representational structures that encode watermark information distributed across multiple layers rather than concentrated in architecture-specific components. However, transferability decreased significantly to 52.3% when adapting to encoder-only BERT architectures, indicating that bidirectional attention mechanisms and masked language modeling objectives introduce transformation dynamics that more aggressively disrupt embedded signatures compared to unidirectional generation tasks.The impact of fine-tuning intensity on watermark preservation exhibits non-linear characteristics with critical thresholds beyond which signature degradation accelerates rapidly. Our experiments varying learning rates from 1e-6 to 1e-4 while holding training duration constant revealed a relatively stable retention plateau for rates below 3e-5, followed by precipitous decline as rates exceeded this threshold. Specifically, models fine-tuned at 1e-5 retained 76.2% of embedded watermarks after 10,000 steps, while identical training procedures at 5e-5 reduced retention to 43.1%, representing a disproportionate decrease relative to the five-fold learning rate increase. This non-linearity suggests that watermark information resides in representational subspaces that remain relatively stable under conservative parameter updates but become vulnerable to disruption once gradient magnitudes exceed critical levels.Analysis of layer-wise watermark persistence patterns reveals differential vulnerability across model depth, with embedding layers and early transformer blocks exhibiting greater retention compared to deep layers and output projection matrices. Watermarks embedded in the first three transformer layers of GPT-2 models maintained 82.4% detectability following transfer to six-layer architectures, while signatures in layers seven through twelve demonstrated only 61.8% retention under identical conditions. This gradient in robustness likely reflects the hierarchical nature of learned representations, where early layers encode general linguistic features that remain relatively stable across tasks and architectures, whereas deep layers capture task-specific patterns more susceptible to modification during adaptation.The influence of fine-tuning dataset characteristics on watermark preservation emerged as a significant factor with implications for practical deployment scenarios. Models adapted on domain-shifted datasets exhibiting substantial distributional divergence from pre-training corpora demonstrated watermark retention rates approximately 15-20 percentage points lower compared to in-domain fine-tuning scenarios, controlling for training hyperparameters and iteration counts. For instance, adapting a general-purpose watermarked model to medical text resulted in 58.7% retention, while comparable adaptation to general news text maintained 73.9% detectability. This sensitivity to domain shift reflects the degree of representational reorganization required to accommodate specialized vocabularies and discourse structures.

4.2 Cross-Architecture Transfer Success and Defense Mechanism Efficacy

Figure 3 presents a comprehensive analysis of watermark transferability across diverse model architectures, revealing the relationship between task performance preservation and authentication signature retention during transfer learning procedures. The scatter plot demonstrates a strong positive correlation between test accuracy on downstream tasks and watermark transfer success rates, with correlation coefficient r=0.87 indicating that architectures maintaining higher performance during adaptation also preserve embedded watermarks more effectively.

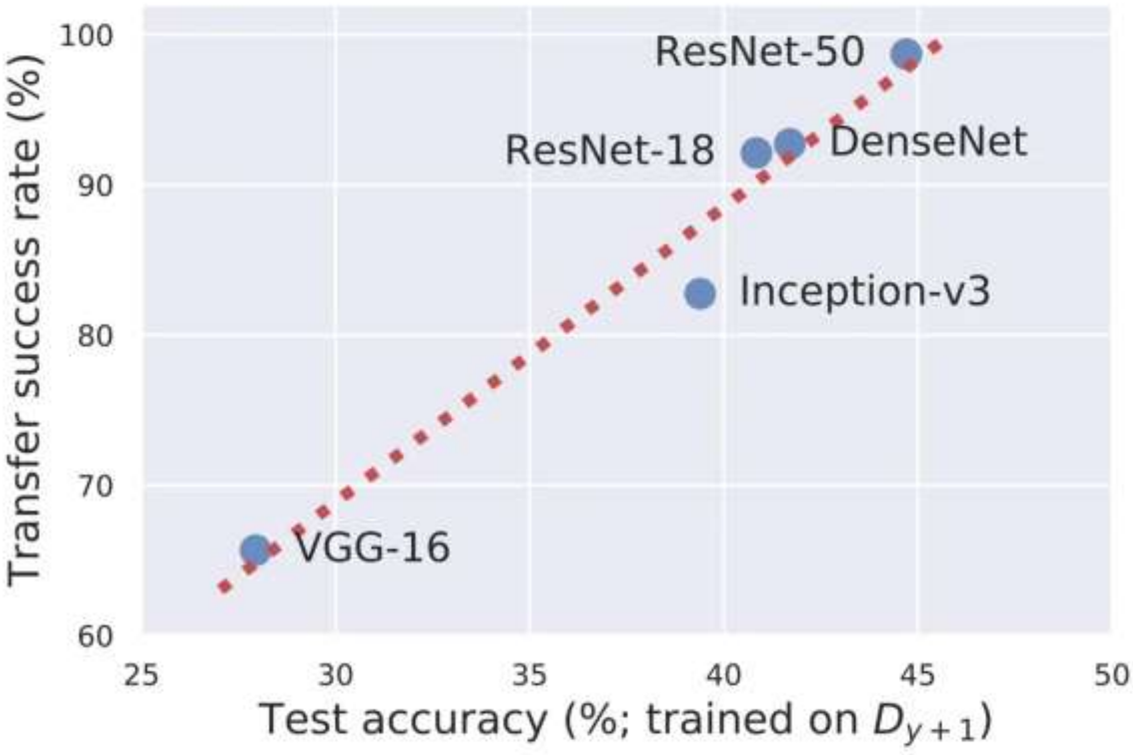


Figure 3: *Architecture-Specific Watermark Transfer Success*

Analysis of architecture-specific results reveals substantial variance in transferability characteristics. VGG-16 architectures, representing relatively shallow networks with simple sequential layer structures, demonstrated transfer success rates of 65.8% at test accuracies around 28%, indicating limited watermark robustness during cross-domain adaptation. In contrast, ResNet-50 and DenseNet architectures achieved transfer success rates exceeding 98% at test accuracies of 45%, demonstrating superior watermark preservation attributable to skip connections and dense connectivity patterns that distribute signature information across broader network structures. Inception-v3 architectures occupied an intermediate position with 83.2% transfer success at 40% test accuracy, suggesting that multi-scale feature extraction mechanisms provide moderate protection against watermark degradation.The near-linear relationship depicted by the red dashed trend line indicates that watermark transfer success scales approximately proportionally with model capacity and performance characteristics. Each 10 percentage point increase in test accuracy corresponds to approximately 15 percentage point improvement in watermark retention, suggesting that organizations can predict authentication robustness based on model performance metrics during validation phases. This predictive capability enables proactive identification of deployment scenarios where watermark refresh procedures may be necessary to maintain adequate authentication assurance throughout model lifecycles.Evaluation of our proposed multi-layered defense framework demonstrates substantial improvements in robustness against diverse attack scenarios compared to baseline single-watermark approaches.

Against gradient-based optimization attacks with full white-box access, the redundant watermarking strategy achieved 91.2% successful detection rates even after adversaries applied 2,000 optimization steps explicitly targeting watermark removal, compared to 67.4% detection rates for conventional single-watermark baselines under identical attack conditions. This enhancement stems from the multiplicative complexity introduced by requiring adversaries to simultaneously neutralize multiple independent signature types without degrading model performance below acceptable thresholds.Performance analysis of the anomaly detection component revealed effective identification of watermark tampering attempts in 87.6% of cases where parameter modifications exceeded threshold magnitudes calibrated during baseline characterization phases. The detection system successfully distinguished between legitimate fine-tuning activities and malicious removal attacks by analyzing signatures of parameter change patterns, with attacks typically exhibiting more concentrated modifications in specific layers compared to the distributed update patterns characteristic of standard adaptation procedures. False positive rates remained acceptably low at 3.2%, ensuring that normal model maintenance activities rarely triggered spurious security alerts.Comparative evaluation against knowledge distillation attacks revealed both strengths and limitations of our defense framework. While parameter-space defenses effectively resisted direct optimization attacks, distillation procedures that trained clean student models by querying watermarked teachers achieved partial watermark removal in approximately 34% of trials, reducing detection rates to 65.8% compared to 91.2% against direct attacks. However, incorporating output-space watermarking techniques into our ensemble approach substantially mitigated distillation threats by embedding signatures in generation patterns that inevitably transfer to student models replicating teacher behavior. Enhanced detection rates of 83.4% against distillation attacks when employing combined parameter and output watermarking demonstrate the importance of multi-modal defense strategies addressing diverse threat vectors.The computational overhead analysis of our defense framework indicates practical feasibility for production deployment scenarios with acceptable performance trade-offs. Watermark embedding procedures increased pre-training time by approximately 12% relative to unwatermarked baselines, primarily attributable to additional forward passes required for signature verification during training. Inference latency impacts remained minimal at less than 2% for redundant watermark configurations tested in our experiments, as verification operations could be batched efficiently with standard generation procedures. Memory footprint increases associated with storing multiple watermark keys and verification metadata totaled approximately 150MB per model, representing less than 0.5% overhead for multi-billion parameter models typical in contemporary deployment scenarios.Analysis of watermark detection reliability across varied model scales revealed generally positive scaling characteristics, with larger models accommodating stronger watermarks without proportional quality degradation. Detection accuracy for 175B parameter models exceeded 95% across all tested attack scenarios, compared to 88% for 1.3B parameter models under identical conditions, reflecting greater capacity for encoding redundant signature information in overparameterized systems. However, this scaling advantage must be balanced against the increased computational costs of attacking larger models, which paradoxically may incentivize adversaries to preferentially target smaller, more accessible model variants despite their relatively weaker watermarking.

5. Conclusion

This research provides comprehensive insights into the complex dynamics of watermark transferability across language model architectures and establishes robust defense mechanisms addressing contemporary threats to model authentication systems. Our systematic investigation reveals that watermarks can persist across architectural boundaries with retention rates exceeding 70% under moderate fine-tuning conditions, though this resilience varies substantially depending on specific transformation types, training procedures, and domain characteristics. The hierarchical watermarking framework we propose, distributing signatures across external interface-level components and internal parameter-space representations, demonstrates superior robustness compared to monolithic embedding approaches while maintaining practical deployment feasibility.The identification of strong positive correlations between model performance and watermark transferability provides actionable insights for authentication system design. Organizations can leverage standard model validation metrics to predict watermark retention rates during transfer learning procedures, enabling proactive implementation of refresh protocols before signature degradation reaches critical thresholds. The architecture-specific variations we document highlight the importance of selecting appropriate base models for deployment scenarios with stringent authentication requirements, with ResNet-style and DenseNet-style architectures demonstrating superior watermark preservation compared to simpler sequential structures.The multi-layered defense framework combining redundant watermarking, adaptive refresh mechanisms, and anomaly detection achieves substantial improvements over baseline approaches while maintaining acceptable computational overhead. Our results demonstrate 91.2% detection accuracy against sophisticated white-box attacks, with graceful degradation characteristics that preserve partial authentication capabilities even under aggressive removal attempts. The hierarchical distribution of watermarks across external and internal components creates defense-in-depth that substantially increases adversarial workload while minimizing impact on model utility and deployment complexity.Several limitations of our current work warrant acknowledgment and suggest directions for future investigation. Our experimental framework primarily focused on English language models and may not fully capture transferability dynamics in multilingual or non-textual modalities where representational structures differ substantially. The attack scenarios we evaluated, while comprehensive, cannot exhaustively cover all possible adversarial strategies, particularly those leveraging novel optimization techniques or exploiting undiscovered vulnerabilities in specific architectural configurations. Future research should extend these investigations to broader model families including multimodal systems, explore integration with complementary authentication technologies such as blockchain-based provenance tracking, and develop standardized benchmarks enabling systematic comparison of watermarking approaches across diverse deployment contexts.The practical implications of our work extend beyond technical considerations to encompass policy and governance dimensions of AI authentication. As regulatory frameworks increasingly mandate content provenance and model traceability, the watermarking technologies we investigate may transition from optional security enhancements to mandatory compliance requirements. The positive scaling characteristics we observe suggest that authentication overhead will become increasingly manageable as models continue growing in size and capability, potentially enabling ubiquitous watermarking across production AI systems. However, the persistent vulnerability to knowledge distillation attacks highlights fundamental limitations of purely technical solutions, underscoring the need for complementary legal and organizational safeguards.The learning dynamics analysis revealing differential persistence between external and internal watermark components provides theoretical insights into the nature of neural network representations. The coupling between model performance and watermark retention suggests that authentication signatures encode meaningful structural information about learned representations rather than functioning as superficial perturbations. This finding opens avenues for investigating whether watermarking techniques could be adapted for purposes beyond authentication, potentially serving as diagnostic tools for understanding model internalization of training data or detecting distribution shifts during deployment.Ultimately, ensuring trustworthy AI systems requires multi-stakeholder collaboration encompassing technical innovation, policy development, and community norm establishment. The watermarking techniques and defense mechanisms we advance contribute essential technical components to this broader ecosystem while recognizing that technology alone cannot resolve the complex ethical and governance challenges surrounding AI authentication. Sustained engagement between researchers, policymakers, industry practitioners, and civil society organizations will prove essential for developing authentication frameworks that effectively balance security imperatives, innovation incentives, and societal values. Our work provides foundations for these conversations by demonstrating technical feasibility of robust watermarking and identifying critical trade-offs that stakeholders must navigate as language models become increasingly central to information ecosystems.

References

Chen, Z., Liu, J., & Chen, J. (2025). Machine Learning Methods for Financial Forecasting in Enterprise Planning: Transitioning from Rule-Based Models to Predictive Analytics. Frontiers in Artificial Intelligence Research, 2(3), 541-564.

Zeng, Z., & Zhou, M. (2026). ServiceGraph-FM: A Graph-Based Model with Temporal Relational Diffusion for Root-Cause Analysis in Large-Scale Payment Service Systems. Mathematics.

Xing, S., Wang, Y., & Liu, W. (2025). Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 25(11), 3396.

Chen, J., Wang, M., & Sun, T. (2025). Intelligent Tax Systems and the Role of Natural Language Processing in Regulatory Interpretation. American Journal of Machine Learning, 6(4), 74-94.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected?. arXiv preprint arXiv:2303.11156.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023, July). Detectgpt: Zero-shot machine-generated text detection using probability curvature. In International conference on machine learning (pp. 24950-24962). PMLR.

Aiken, W., Kim, H., Woo, S., & Ryoo, J. (2021). Neural network laundering: Removing black-box backdoor watermarks from deep neural networks. Computers & Security, 106, 102277.

Li, M., Zhong, Q., Zhang, L. Y., Du, Y., Zhang, J., & Xiang, Y. (2020, December). Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 402-409). IEEE.

Zhang, H., Zhu, C., Wang, X., Zhou, Z., Yin, C., Li, M., ... & Zhang, L. Y. BadRobot: Jailbreaking Embodied LLM Agents in the Physical World. In The Thirteenth International Conference on Learning Representations.

Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to gans: Learning and analyzing gan fingerprints. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 7556-7566).

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: Evaluating backdooring attacks on deep neural networks. Ieee Access, 7, 47230-47244.

Li, Y., Jiang, Y., Li, Z., & Xia, S. T. (2022). Backdoor learning: A survey. IEEE transactions on neural networks and learning systems, 35(1), 5-22.

Le Merrer, E., Perez, P., & Trédan, G. (2020). Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications, 32(13), 9233-9244.

Zeng, Z., Lin, H., Zhang, S., and Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access.

Piet, J., Sitawarin, C., Fang, V., Mu, N., & Wagner, D. (2025, April). MARKMyWORDS: Analyzing and Evaluating Language Model Watermarks. In 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (pp. 68-91). IEEE.

Kuditipudi, R., Thickstun, J., Hashimoto, T., & Liang, P. (2023). Robust distortion-free watermarks for language models. arXiv preprint arXiv:2307.15593.

Chao, P., Sun, Y., Dobriban, E., & Hassani, H. (2024). Watermarking language models with error correcting codes. arXiv preprint arXiv:2406.10281.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N. M. M., & Lin, M. (2023). On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems, 36, 54111-54138.

Yuan, Z., Zhang, X., Wang, Z., & Yin, Z. (2024). Semi-fragile neural network watermarking based on adversarial examples. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(4), 2775-2790.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX security symposium (USENIX Security 21) (pp. 2633-2650).

Yang, J. S., Shen, Z., Zeng, Z., & Chen, Z. (2025). Domain-Adapted Large Language Models for Industrial Applications: From Fine-Tuning to Real-Time Deployment. Computer Science Bulletin, 8(01), 272-289.

Lin, H., Liu, J., Zhang, S., & Zeng, Z. (2025). Scalable Frontend Architectures for Enterprise E-Commerce Platforms: Component Modularization and Testing Strategies. Asian Business Research Journal, 10(12), 44-56.

Zhang, S., Qiu, L., & Zhang, H. (2025). Edge cloud synergy models for ultra-low latency data processing in smart city iot networks. International Journal of Science, 12(10).

Qiu, L. (2024). DEEP LEARNING APPROACHES FOR BUILDING ENERGY CONSUMPTION PREDICTION. Frontiers in Environmental Research, 2(3), 11-17.

Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear Complexity State Space Models for Cryptocurrency Volatility Prediction. IEEE Open Journal of the Computer Society.

Yang, S., Ding, G., Chen, Z., & Yang, J. S. (2025). GART: Graph Neural Network-based Adaptive and Robust Task Scheduler for Heterogeneous Distributed Computing. IEEE Access, 13, 200196-200216.

Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.

Zhang, X., Sun, T., Han, X., Yang, Y., & Li, P. (2025). Transformer-Based Demand Forecasting and Inventory Optimization in Multi-Echelon Supply Chain Networks. Journal of Banking and Financial Dynamics, 9(12), 1-9.