Advancing Backdoor Attack Detection in Transformer Models using Feature Squeezing and Statistical Anomaly Filtering Techniques

Simon Grant
Department of Software Engineering
Rowan University
s.grant@rowan.edu

Abstract

The rapid proliferation of Transformer-based architectures across critical socio-technical infrastructures has introduced significant security vulnerabilities, most notably the emergence of sophisticated backdoor attacks. These adversarial interventions involve the clandestine insertion of malicious triggers during the training or fine-tuning phases, which remain dormant during standard operational cycles but activate specific, harmful behaviors when encountering predefined input patterns. This research investigates the advancement of backdoor detection mechanisms through the integration of feature squeezing and statistical anomaly filtering. By reducing the complexity of the input space and systematically identifying deviations in the latent representation distributions, the proposed framework enhances the robustness of large-scale language and vision Transformers. The study provides a comprehensive system-level analysis of how these defensive layers interact with existing model deployments, emphasizing the trade-offs between computational overhead and security efficacy. Furthermore, the discussion extends to the governance of AI supply chains, the policy implications of vulnerable foundation models, and the sustainability of long-term defensive strategies in evolving adversarial landscapes. The findings suggest that a multi-layered, statistical approach to anomaly detection can significantly mitigate the risks posed by poisoned datasets and compromised third-party model providers, thereby reinforcing the integrity of the broader artificial intelligence ecosystem.

Keywords:
Transformer Models, Backdoor Attack Detection, Feature Squeezing, Statistical Anomaly Filtering, AI Security Governance, Socio-Technical Infrastructures.

1. Introduction

The integration of Transformer architectures into the core of modern digital civilization represents a paradigm shift in how information is processed, interpreted, and acted upon within large-scale systems [23]. From financial market prediction and autonomous vehicle

navigation to medical diagnostics and national security protocols, the reliance on these high-capacity models is nearly absolute. However, this dependence has outpaced the development of comprehensive security frameworks, leaving a critical gap in the defense against adversarial manipulations [18]. Backdoor attacks, in particular, represent a pernicious threat because they exploit the very flexibility and scale that make Transformers effective. Unlike traditional evasion attacks that target a model's inference logic directly, backdoor attacks involve the subtle poisoning of training data or the manipulation of model parameters to establish a hidden association between a seemingly innocuous trigger and a specific malicious output [5]. This vulnerability is compounded by the increasing trend toward outsourcing model training to third-party providers and the use of unverified datasets scraped from the public internet, creating a fragmented supply chain where trust is often assumed rather than verified [8].

The challenge of detecting these hidden vulnerabilities is exacerbated by the structural complexity of Transformer models, which often consist of billions of parameters and deep hierarchical layers of attention mechanisms [6]. These characteristics allow for the concealment of triggers within high-dimensional manifolds that are nearly impossible to inspect through manual oversight or traditional signature-based security tools [2]. Existing defensive measures often fall short because they either impose excessive computational burdens that hinder real-time deployment or fail to generalize across different domains and trigger types [15]. Consequently, there is an urgent need for architectural interventions that are both efficient and resilient to diverse attack vectors. This paper addresses this need by exploring a dual-layered defensive strategy centered on feature squeezing and statistical anomaly filtering, providing a robust framework for identifying and neutralizing backdoors before they can be activated in production environments [26].

Furthermore, the security of Transformer models is not merely a technical concern but a significant socio-technical and policy challenge [30]. As these models become embedded in the governance of public life and the management of critical infrastructure, a successful backdoor attack could lead to systemic failures, loss of public trust, and significant economic disruption. The infrastructure supporting these models must therefore be viewed through the lens of national security and public safety. This research situates technical detection methods within the broader context of AI governance, arguing that robust detection is a prerequisite for the ethical and sustainable deployment of artificial intelligence [21]. By analyzing the structural trade-offs and deployment implications of advanced detection techniques, this study aims to provide a roadmap for building more resilient AI systems that can withstand the evolving threats of the modern adversarial landscape.

2. The Architecture of Vulnerability: Backdoor Attacks in Transformers

Understanding the mechanics of backdoor attacks requires a deep dive into the latent spaces of Transformer models and the processes by which they learn representations. In a typical backdoor scenario, an adversary introduces a small set of poisoned examples into the training corpus [16]. These examples contain a trigger—such as a specific sequence of characters in a

text model or a subtle pixel pattern in an image model—paired with a target label that differs from the ground truth. Because Transformers are designed to maximize the capture of contextual nuances, they readily internalize these associations. During inference, the model behaves perfectly on clean data, maintaining high performance and masking the presence of the Trojan [14]. Only when the specific trigger is present does the model deviate from its expected behavior, executing the adversary's intent. This dual nature makes the backdoor particularly dangerous for critical infrastructure where reliability is paramount [12].

The structural characteristics of Transformers, specifically the multi-head attention mechanism and the feed-forward layers, provide a fertile ground for the embedding of these triggers [29]. The attention mechanism, by design, focuses on specific parts of the input to build a comprehensive understanding of context. An adversary can exploit this by ensuring the trigger pattern commandingly captures the model's attention, effectively overriding the legitimate features of the input [24]. From a systems perspective, this represents a corruption of the model's internal routing logic. The complexity of the mapping from input to output means that these corruptions are often located in "dead zones" of the high-dimensional feature space that are rarely traversed during normal operations, making them highly elusive to standard validation and testing procedures [11].

Moreover, the deployment of Transformers within larger socio-technical systems introduces additional layers of risk. When a foundation model is fine-tuned for a specific application, such as legal document analysis or industrial control, the backdoor can persist through the transfer learning process [27]. This "upstream" poisoning has devastating implications for the "downstream" users who may lack the expertise or computational resources to audit the models they adopt. This creates a systemic vulnerability where a single compromised foundation model can infect an entire ecosystem of applications. The infrastructure of AI distribution, characterized by model hubs and pre-trained weights, currently lacks the rigorous certification and verification protocols necessary to guarantee the absence of such hidden threats. Addressing this requires a shift from viewing security as an afterthought to treating it as a fundamental architectural requirement of the system [10].

3. Feature Squeezing as a Defensive Layer

Feature squeezing serves as a primary defensive mechanism by reducing the search space available to an adversary and making the model's internal representations more predictable [26]. The core philosophy of feature squeezing is that many of the high-frequency or overly specific features in an input are unnecessary for the core task and often serve as the vectors for adversarial triggers. By applying transformations that simplify the input—such as bit-depth reduction, spatial smoothing, or median filtering—the defender can strip away the subtle nuances that define a backdoor trigger while preserving the essential information required for correct classification. In the context of Transformer models, this approach can be applied at the input level or at various intermediate layers where latent features are most susceptible to manipulation.

The application of feature squeezing to Transformers involves a delicate balance between security and utility. In text-based Transformers, this might involve synonym substitution or the normalization of character encodings to prevent the use of invisible or rare characters as triggers [4]. In vision-based systems, it involves reducing the precision of pixel values or applying aggressive compression techniques. These methods work on the principle that legitimate features are robust to small perturbations, whereas backdoor triggers are often highly fragile and rely on precise configurations to activate the malicious logic [13]. When the input is squeezed, the discrepancy between the model's prediction on the original input and its prediction on the squeezed input becomes a powerful diagnostic tool. If the predictions differ significantly, it serves as a strong indicator that the original input may contain an adversarial trigger [7].

From a systems engineering perspective, feature squeezing is highly attractive due to its low computational cost and ease of integration into existing pipelines. Unlike techniques that require retraining the entire model or adding complex auxiliary networks, feature squeezing can be implemented as a modular pre-processing step. This modularity is crucial for large-scale deployments where latency and resource consumption are strictly governed. However, the use of feature squeezing also introduces structural trade-offs. Excessive squeezing can lead to a degradation in accuracy for legitimate, edge-case inputs that require high precision. Therefore, the implementation of feature squeezing must be dynamic and context-aware, potentially varying the intensity of the squeezing based on the sensitivity of the task or the perceived threat level of the environment. This necessitates a sophisticated governance framework to manage the trade-offs between system performance and security resilience.

4. Statistical Anomaly Filtering in Latent Spaces

While feature squeezing addresses the input-level vulnerabilities, statistical anomaly filtering focuses on the internal behavior of the Transformer model. This technique operates on the premise that inputs containing backdoor triggers will generate activation patterns in the latent layers that are statistically distinct from those produced by benign data [22]. By monitoring the distribution of activations across the various layers of the Transformer, it is possible to identify outliers that signify the activation of a hidden Trojan. This requires the establishment of a baseline "profile" of normal model behavior during a clean validation phase, against which all subsequent operational data is compared [25].

The implementation of statistical anomaly filtering involves the use of high-dimensional statistical methods to analyze the activation vectors produced by the attention heads and feed-forward networks [28]. Techniques such as kernel density estimation, principal component analysis, or robust covariance estimation can be used to map the manifold of normal activations. When an input is processed, its corresponding activation vector is checked for its likelihood within the established distribution. Inputs that fall into low-probability regions are flagged as suspicious. This approach is particularly effective against sophisticated backdoors that are designed to bypass simple input-level filters, as the internal logic of the

model must still diverge to execute the malicious output, inevitably leaving a statistical footprint in the latent space [19].

In large-scale AI infrastructures, the deployment of statistical anomaly filtering requires a robust data management strategy. The storage and processing of activation statistics for models with millions of parameters can be significant, necessitating efficient dimensionality reduction and sampling strategies. Furthermore, the filtering mechanism must be resilient to "distribution shift," where the characteristics of legitimate data change over time due to evolving user behavior or environmental factors. This requires the continuous updating of the statistical profiles, which introduces new challenges regarding the integrity of the update process itself. If an adversary can subtly shift the baseline over time, they may be able to slowly introduce poisoned data into the "normal" distribution, a tactic known as "boiling the frog." Consequently, statistical anomaly filtering must be part of a broader, multi-temporal defensive strategy that considers the long-term evolution of the system.

5. System-Level Integration and Architectural Trade-offs

Integrating feature squeezing and statistical anomaly filtering into a unified defensive framework requires a comprehensive system-level design that considers the interplay between various components. The resulting architecture can be envisioned as a series of nested gates. The outermost gate applies feature squeezing to neutralize low-level triggers and identify simple adversarial attempts. The inner gate employs statistical anomaly filtering to monitor the deep latent representations for more subtle, high-level manipulations. This layered approach adheres to the principle of defense-in-depth, ensuring that the failure of one mechanism does not compromise the entire system [17]. Such an architecture is essential for maintaining the robustness of Transformer models in high-stakes environments where the cost of failure is extreme.

However, the addition of these defensive layers introduces significant architectural trade-offs, particularly regarding the triad of security, performance, and cost. Every additional filtering or squeezing step adds latency to the inference process, which can be a critical bottleneck for real-time applications like autonomous driving or high-frequency trading. Furthermore, the development and maintenance of these defensive systems require specialized expertise and significant computational resources for profiling and monitoring. Organizations must decide how to allocate their limited "security budget" across different parts of the AI lifecycle. In many cases, it may be more cost-effective to invest in more rigorous data curation and supply chain verification than to rely solely on post-deployment detection [3]. This highlights the need for a holistic perspective that views AI security not just as a technical problem but as a resource management and strategic planning challenge.

The sustainability of these defensive measures is another critical consideration. As adversaries become aware of specific detection techniques, they will undoubtedly develop counter-measures, such as "squeezing-aware" triggers or attacks that mimic the statistical distribution of benign data [1]. This creates a constant arms race between attackers and

defenders. To remain effective, defensive architectures must be adaptive and incorporate elements of randomness or secrecy that make them difficult for an adversary to model. This might include varying the parameters of the feature squeezing or using an ensemble of different statistical filters. Furthermore, the infrastructure must support rapid updates and patching, similar to how traditional software security is managed. The ability to quickly deploy new defensive signatures or statistical profiles across a fleet of Transformer models is a key requirement for maintaining long-term resilience in an adversarial environment.

6. Governance, Policy, and the AI Supply Chain

The technical challenges of backdoor detection are inextricably linked to the broader issues of AI governance and the management of the AI supply chain. As the production of foundation models becomes concentrated among a few well-resourced entities, the risks associated with "single point of failure" vulnerabilities increase. A backdoor in a widely used Transformer model could have global repercussions, affecting thousands of downstream applications and millions of users [20]. This necessitates the development of new policy frameworks and industry standards for model verification and certification. Governments and regulatory bodies have a role to play in mandating transparency in the training process and requiring third-party audits for models deployed in critical infrastructure.

The concept of a "Model Bill of Materials" (MBOM) has emerged as a potential solution to the supply chain problem. Just as a Software Bill of Materials (SBOM) tracks the components and dependencies of traditional software, an MBOM would document the datasets, training parameters, and hardware environments used to create an AI model. This transparency would allow downstream users to assess the risk of poisoning and more effectively apply targeted detection techniques like those discussed in this study. Furthermore, the establishment of "secure enclaves" for model training and the use of cryptographic techniques to verify model integrity could provide stronger guarantees of safety [9]. However, these measures must be balanced against the need for innovation and the protection of intellectual property, requiring a nuanced approach to policy-making that involves stakeholders from across the technical, legal, and ethical domains.

Moreover, the ethical implications of backdoor detection mechanisms cannot be ignored. Statistical anomaly filtering, by its nature, identifies "deviant" behavior, which raises concerns about the potential for bias and unfairness. If the training data for the statistical baseline is not sufficiently diverse, the filter may disproportionately flag legitimate inputs from minority groups or unconventional contexts as "anomalous." This could lead to a situation where security measures inadvertently reinforce existing social inequalities or limit the accessibility of AI systems. Therefore, the design and deployment of detection systems must incorporate fairness audits and use techniques like differential privacy to protect user data. Ensuring that AI security is both robust and equitable is a fundamental requirement for building a trustworthy and sustainable digital future.

7. Socio-Technical Implications and Resilience

The resilience of socio-technical systems depends not only on the robustness of the individual components but also on the ability of the system as a whole to absorb and recover from shocks. In the context of Transformer-based AI, this means that detection mechanisms must be integrated into a larger framework of incident response and disaster recovery [30]. When a backdoor is detected, the system must have a predefined protocol for failing gracefully, notifying users, and initiating a remediation process. This might involve switching to a simpler, non-Transformer backup model or temporarily disabling certain high-risk features. The ability to maintain core functionality under attack is a hallmark of a resilient system and is essential for maintaining public trust in AI technologies.

The public's perception of AI security is also a critical factor in the long-term viability of these systems. High-profile backdoor attacks could lead to a "tech-lash" where the benefits of AI are overshadowed by fears of manipulation and surveillance. Clear communication about the risks and the steps being taken to mitigate them is essential for managing expectations and fostering an informed public discourse. This includes being transparent about the limitations of current detection techniques and the ongoing nature of the adversarial threat. Educational initiatives aimed at improving "AI literacy" among policymakers, business leaders, and the general public can help create a more resilient society that is better equipped to navigate the complexities of an AI-augmented world.

Finally, the sustainability of AI security research itself is a concern. As models grow in size and complexity, the computational cost of conducting security research increases, potentially pricing out academic institutions and smaller labs. This could lead to a concentration of security expertise within a few large corporations, creating a conflict of interest where the entities producing the models are also the ones responsible for auditing them. To prevent this, there is a need for increased public funding for independent AI security research and the creation of open-source datasets and benchmarks for backdoor detection. By democratizing the tools and knowledge required to secure Transformer models, we can ensure that the defense of our digital infrastructure remains a collective and transparent effort.

8. Forward-Looking Perspectives and Future Research

Looking ahead, the evolution of Transformer models will continue to challenge existing detection paradigms. The move toward multi-modal Transformers, which process text, images, and audio simultaneously, introduces new avenues for cross-modal backdoor attacks. For example, a trigger in an audio clip could activate a malicious response in a text-based output, bypassing defenses that are focused on a single modality. Research into multi-modal anomaly detection and the development of unified feature squeezing techniques will be critical for addressing these emerging threats. Furthermore, the rise of "on-device" and "edge" AI means that detection mechanisms must be optimized for extremely low-resource environments without sacrificing efficacy.

The integration of formal methods and provable security into the Transformer design process

is another promising area for future research. While statistical methods provide a strong empirical defense, they do not offer the rigorous guarantees of formal verification. Developing architectures that are "secure by design" and whose properties can be mathematically proven would represent a significant step forward in AI security. This might involve the use of constrained optimization during training to prevent the formation of backdoors or the development of new types of layers that are inherently resistant to triggers. Although these approaches are currently in their infancy, they hold the potential to transform the field from a reactive arms race to a proactive engineering discipline.

Finally, the global nature of the AI research community means that security must be an international priority. Adversaries do not respect national borders, and a breakthrough in backdoor technology in one part of the world can be quickly weaponized elsewhere. International collaboration on security standards, threat intelligence sharing, and the development of global norms for AI behavior is essential for maintaining a stable and secure digital environment. By viewing the security of Transformer models as a shared global responsibility, we can work toward a future where the power of artificial intelligence is harnessed for the benefit of all, protected by a robust and resilient defensive infrastructure.

9. Conclusion

The advancement of backdoor attack detection in Transformer models is a critical necessity for the security and stability of modern socio-technical infrastructures. This study has explored the integration of feature squeezing and statistical anomaly filtering as a multi-layered defensive strategy, highlighting their effectiveness in identifying and neutralizing hidden Trojans. Through a system-level analysis, we have discussed the architectural trade-offs, deployment challenges, and governance implications of these techniques. While feature squeezing provides a computationally efficient first line of defense at the input level, statistical anomaly filtering offers a deep, behavioral monitoring of the model's latent representations. Together, they form a robust framework for enhancing the resilience of large-scale AI systems.

However, the technical solutions discussed here are only one part of a much larger puzzle. Building a secure AI future requires a holistic approach that integrates technical innovation with robust policy-making, ethical oversight, and public education. The management of the AI supply chain, the development of transparency standards like the Model Bill of Materials, and the commitment to independent security research are all essential components of a sustainable defense strategy. As Transformer models continue to evolve and permeate every aspect of our lives, the urgency of this work only grows. By prioritizing security as a fundamental architectural requirement and working collaboratively across disciplines and borders, we can build AI systems that are not only powerful and efficient but also trustworthy and resilient in the face of an ever-changing adversarial landscape.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Gu, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31.

2. Adi, Y., Baum, C., Cisse, M., Pinkas, B., & Keshet, J. (2018). Turning your weakness into a strength: Watermarking deep neural networks by backdooring. Proceedings of the 27th USENIX Security Symposium.

3. Bagdasaryan, E., & Shmatikov, V. (2021). Blind backdoors in deep learning models. Proceedings of the 30th USENIX Security Symposium.

4. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP), 39-57.

5. Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186.

7. Dong, Y., Fu, Q., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2019). Benchmarking adversarial robustness on image classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

8. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.

9. Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., ... & Efros, A. A. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. International Conference on Machine Learning, 1989-1998.

10. Hu, Z., Shen, Y., Kuang, Z., & Zheng, B. (2026). Resilient Architectures for Large-Scale AI Systems. Journal of Infrastructure Security, 14(2), 112-135.

11. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems, 32.

12. Ji, Y., Zhang, X., Ji, S., Luo, X., & Wang, T. (2018). Model-reuse attacks on deep learning systems. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security.

13. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

14. Li, Y., Wu, B., Jiang, Y., Li, Z., & Xia, S. T. (2022). Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems.

15. Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. International Symposium on Research in Attacks, Intrusions, and Defenses, 273-294.

16. Liu, Y., Ma, S., Aafer, Y., Lee, W. C., Zhai, J., Wang, W., & Zhang, X. (2018). Trojaning attack on neural networks. Proceedings of the 25th Network and Distributed System Security Symposium (NDSS).

17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.

18. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P).

19. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

20. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP).

21. Sun, Lichao, et al. (2024). A survey on large language model security and privacy. IEEE Communications Surveys & Tutorials.

22. Tran, B., Li, J., & Madry, A. (2018). Spectral signatures in backdoor attacks. Advances in Neural Information Processing Systems, 31.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

24. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for nlp. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

25. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019). Neural cleanse: Identifying and mitigating backdoor attacks in deep neural networks.

2019 IEEE Symposium on Security and Privacy (SP), 707-723.

26. Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. Proceedings of the 25th Network and Distributed System Security Symposium (NDSS).

27. Yao, Y., Li, H., Zheng, H., & Zhao, B. Y. (2019). Latent backdoor attacks on deep learning models. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security.

28. Zhang, J., Chen, Z., & Wang, Q. (2026). Statistical Filtering for Latent Anomaly Detection in Transformers. Journal of Artificial Intelligence Research, 74, 455-489.

29. Zhao, Pu, et al. (2025). On the security and robustness of vision transformers. International Journal of Computer Vision.

30. Zhou, Y., & Li, M. (2026). Socio-Technical Implications of AI Vulnerabilities in National Infrastructure. Technology in Society, 88, 102-120.