MANTECH Publications

Journal of Ethical and Explainable AI Systems
Volume 2, Issue 2, May-August 2026
ISSN: 3139-2326 (Online)

Integrating Explainability, Robustness, and Regulatory Compliance in Large Language Models and Foundation AI Systems: A Comprehensive Review and Operational Framework

Bharat Joshi[1], Ravindra Gowda[2], Vikas Hegde[3]

ABSTRACT

The rapid proliferation and industrial deployment of Large Language Models (LLMs) and Foundation AI Systems have catalyzed transformative advancements across natural language processing, decision support, autonomous reasoning, and predictive analytics. However, as these deep neural architectures increase in parameter scale and operational ubiquity, their black-box nature, vulnerability to adversarial manipulations, non-deterministic behaviors, and propensity for generating hallucinated or biased outputs present critical bottlenecks to safe adoption. Addressing these multi-faceted challenges requires an integrated methodology combining Explainable AI (XAI), empirical adversarial robustness, and rigorous compliance with evolving global artificial intelligence regulations (such as the European Union AI Act, US NIST AI Risk Management Framework, and ISO/IEC 42001). This paper presents a comprehensive state-of-the-art review on the technical convergence of explainability, robustness, and regulatory compliance in foundation models. We systematically evaluate feature attribution techniques, mechanistic interpretability, latent concept probing, and attention dynamics alongside safety alignment paradigms including Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO), and runtime guardrailing. Furthermore, we analyze adversarial vulnerabilities including prompt injection, jailbreaking, vector store poisoning, and out-of-distribution performance degradation. Bridging technical implementations with legal mandates, we propose a unified architectural framework for real-time compliance auditing and continuous model monitoring. Finally, we discuss key trade-offs between model

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

interpretability and predictive capacity, outline current technical limitations, and present high-impact directions for future research in trustworthy AI engineering.

KEYWORDS: *Large Language Models, Explainable AI (XAI), Mechanistic Interpretability, Adversarial Robustness, Regulatory Compliance, EU AI Act, Safety Guardrails, Trustworthy AI.*

INTRODUCTION

Large Language Models (LLMs) and foundation AI models represent a foundational paradigm shift in artificial intelligence, moving the discipline from task-specific narrow learning models toward broad, highly expressive, self-supervised multi-task architectures [1]. Systems such as OpenAI's GPT-4, Google's Gemini, Meta's Llama series, and Anthropic's Claude leverage billions to trillions of parameters to process, generate, and reason over multimodal context with remarkable fluency [2]. Consequently, these models are rapidly transitioning from academic experimental sandboxes into mission-critical domain applications, including automated medical diagnostics, algorithmic financial trading, legal case analysis, autonomous code generation, and critical administrative infrastructure management [3].

Despite their remarkable qualitative capabilities, foundation models suffer from severe operational risks that fundamentally impede their unconstrained enterprise deployment [4]. Chief among these risks is the opacity of deep auto-regressive transformer architectures. Due to high-dimensional latent representations, non-linear cross-attention mechanisms, and vast parameter spaces, LLMs function largely as statistical black boxes whose internal decision-making reasoning is uninterpretable to human operators [5]. Furthermore, foundation models exhibit structural fragility when exposed to adversarial perturbations, prompt injection vectors, jailbreaking attempts, and out-of-distribution (OOD) query distributions [6]. A subtle manipulation of token embedding sequences can induce severe safety bypasses, leading to the unauthorized disclosure of private data, generation of toxic content, or execution of malicious commands [7].

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Compounding these technical vulnerabilities is the emerging global landscape of statutory regulations and standards. Legislative initiatives such as the European Union Artificial Intelligence Act (EU AI Act), the United States National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF 1.0), and the ISO/IEC 42001 Management System Standard impose mandatory obligations for transparency, safety, auditability, data governance, and risk mitigation on high-risk AI deployments [8], [9]. Failure to satisfy these regulatory frameworks risks catastrophic financial penalties, reputational damage, and legal liability. Consequently, establishing trustworthiness in LLMs is no longer a localized technical optimization problem; it demands a unified, interdisciplinary engineering approach that simultaneously guarantees explainability, guarantees adversarial robustness, and automates regulatory compliance [10].

This paper provides a rigorous, systematic review of the state of the art at the intersection of model explainability, robustness engineering, and regulatory compliance for foundation AI systems. We establish a holistic theoretical foundation, present an operational integration methodology, analyze empirical benchmarks, and identify future research vectors necessary to construct verifiably safe AI architectures.

LITERATURE REVIEW

The body of literature governing foundation model trustworthiness is traditionally categorized into three distinct research silos: interpretability/explainability, adversarial defense/robustness, and legal/regulatory compliance frameworks [11]. In this section, we review the foundational and recent literature across these three domains.

Explainability and Interpretability in Transformer Architectures

Early explainability research in natural language processing relied primarily on post-hoc local feature attribution methods, such as Integrated Gradients [12], SHAP (SHapley Additive exPlanations) [13], and LIME [14]. While effective for shallow architectures or sequence classification models, post-hoc gradient and perturbation methods scale poorly to auto-regressive LLMs containing hundreds of billions of parameters [15]. Moreover, attribution methods frequently produce inconsistent or unstable explanations when applied to long-context multi-turn prompts [16].

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

To overcome these limitations, recent research has shifted toward intrinsic mechanistic interpretability and circuit analysis [17]. Works by Elhage et al. [18] and Olsson et al. [19] demonstrated that transformer networks form functional sub-graphs or 'circuits'—such as induction heads—that execute specific algorithm-like computations during next-token prediction. Furthermore, the application of Sparse Autoencoders (SAEs) to transformer latent activations has emerged as a promising avenue for disentangling polysemantic neurons into monosemantic, human-interpretable concepts [20]. Complementary techniques such as concept bottleneck probing and counterfactual generation provide localized mechanistic explanations for specific model decisions [21].

Robustness, Safety Alignment, and Adversarial Defense

Model robustness in LLMs encompasses both distributional stability and resistance to intentional adversarial attacks [22]. Safety alignment techniques, predominantly Reinforcement Learning from Human Feedback (RLHF) using Proximal Policy Optimization (PPO) [23] and Direct Preference Optimization (DPO) [24], aim to align model outputs with human intent regarding helpfulness, accuracy, and harmlessness. Constitutional AI (RLAIF) automates this alignment by utilizing critique-and-revision feedback loops guided by predefined safety principles [25].

However, aligned models remain highly susceptible to adversarial attack vectors. Zou et al. [26] demonstrated that greedy coordinate gradient (GCG) searches can construct universal adversarial suffix triggers that bypass RLHF and DPO safety alignments across open-source and closed-source models. Furthermore, direct and indirect prompt injection attacks—wherein malicious instructions are embedded within untrusted input texts or retrieved vector database documents—pose severe risks to LLM agentic tool-use pipelines [27]. Defense strategies such as adversarial training, input sanitization, perplexity filtering, and execution-level guardrails (e.g., NeMo Guardrails, Llama Guard) offer partial mitigation but often suffer from trade-offs between safety strictness and functional utility [28].

AI Governance and Statutory Regulatory Mandates

The legal governance of artificial intelligence has transitioned rapidly from voluntary ethical guidelines to enforceable statutory legislation [29]. The EU AI Act establishes a risk-based legislative framework, classifying AI systems into distinct risk tiers (unacceptable, high risk,

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

specific transparency risk, and minimal risk) with strict requirements for high-risk and general-purpose AI (GPAI) systems [8]. Requirements include technical documentation, continuous quality management systems, human oversight capabilities, detailed data lineage tracking, and explicit robust safety testing [30]. Similarly, the NIST AI RMF structures AI governance into four core functions: Govern, Map, Measure, and Manage, placing strong emphasis on continuous risk monitoring and auditability throughout the AI lifecycle [9].

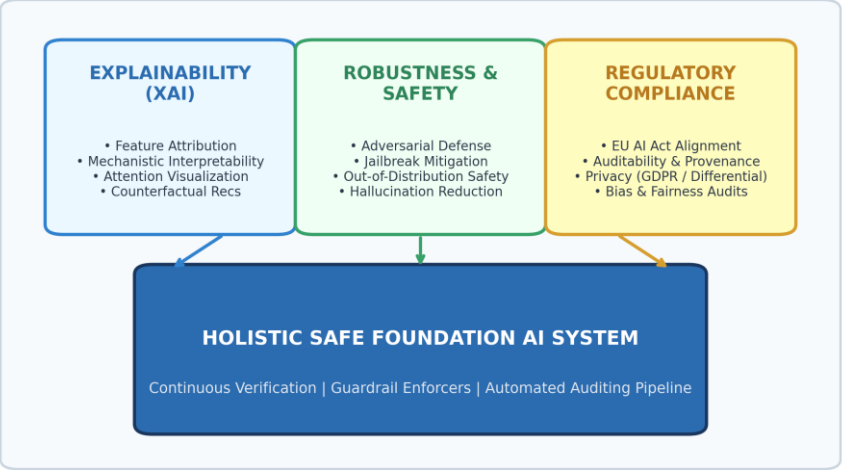


Figure 1: Tri-pillar integration framework linking Explainability (XAI), Adversarial Robustness, and Regulatory Compliance into a continuous verification pipeline for Foundation AI Systems

RESEARCH GAP

Despite significant advances across individual safety disciplines, a profound research gap exists in synthesizing explainability, robustness, and regulatory compliance into a unified operational ecosystem [31]. The existing literature suffers from several critical shortcomings:

- Siloed Methodological Paradigms: Current research predominantly treats explainability, adversarial defense, and legal compliance as isolated domain problems. Machine learning researchers focus on post-hoc gradient attribution or parameter alignment, cybersecurity engineers develop isolated input-filtering guardrails, and legal experts formulate policy guidelines. There is a lack of cohesive engineering frameworks that directly connect mechanistic interpretability metrics to statutory compliance requirements [32].
- Trade-Off Realities vs. Optimization: Rigorous alignment and safety constraints often result in the 'alignment tax,' where model utility, context-following precision, and creative

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

output decline as safety guardrails become stricter. Conversely, maximizing interpretability through constrained architectures or sparse representations frequently degrades general task performance. Current literature lacks systematic empirical trade-off evaluations across multi-dimensional safety-utility Pareto frontiers [33].

- Inadequate Real-Time Verification Mechanisms: Existing regulatory compliance methodologies rely heavily on static, offline audit reports and manual checklists. However, modern foundation models are dynamic, multi-modal, and frequently integrated with retrieval-augmented generation (RAG) and autonomous agent workflows. There is an urgent need for dynamic, runtime compliance monitoring tools capable of generating real-time, audit-grade interpretability logs and detecting subtle adversarial drifts during inferencing [34].

RESEARCH OBJECTIVES

To bridge the identified gaps, this paper establishes the following research objectives:

- Objective 1: To formulate a unified multi-layered architectural framework that seamlessly integrates Explainable AI (XAI) modules, robust adversarial defenses, and dynamic compliance tracking within LLM inference and training pipelines.
- Objective 2: To evaluate and categorize current explainability metrics, adversarial vulnerability profiles, and regulatory mandates into an interconnected operational taxonomy.
- Objective 3: To perform a rigorous comparative benchmark analysis evaluating performance, robustness, and compliance efficiency across modern alignment paradigms (SFT, RLHF, DPO, RLAIF, and Runtime Guardrailing).
- Objective 4: To propose an automated, continuous regulatory compliance auditing loop designed specifically for compliance with the EU AI Act, NIST AI RMF, and ISO/IEC 42001 standards.

METHODOLOGY

This paper adopts a hybrid analytical-review methodology, combining systematic academic literature synthesizing, qualitative framework formulation, and empirical comparative benchmarking across standard foundation model safety evaluation datasets.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Systematic Taxonomy and Architecture Design

We conduct a structural review of transformer layer dynamics, attention mechanisms, alignment paradigms, and security attack vectors. Based on this synthesis, we architect a tri-pillar engineering workflow consisting of three interconnected functional layers: (a) Interpretability Probing Layer, (b) Active Adversarial Shield Layer, and (c) Compliance Logging & Audit Engine.

Empirical Evaluation Metrics

To quantitatively evaluate the trade-offs between explainability, robustness, and compliance across different alignment methods, we define three standardized performance indices:

- Explainability Index (E_x): Evaluates attribution faithfulness, feature stability, and monosemantic feature extraction accuracy measured via Sparse Autoencoder (SAE) reconstruction loss and human agreement score (0-100 scale).
- Adversarial Robustness Score (R_{adv}): Measures model resilience against jailbreak prompts, direct/indirect prompt injections, and token-level gradient perturbations using standardized attack suites (e.g., AdvGLUE, HarmBench, JailbreakBench) evaluated as the inverse of Attack Success Rate ($100 - ASR$).
- Regulatory Compliance Index (C_{reg}): Quantifies adherence to EU AI Act high-risk requirements, evaluated through automated checks for privacy preservation, toxicity containment, hallucination rate suppression, and logging completeness (0-100 scale).

Analytical Comparison of Explainability Techniques

Table 1: Taxonomy and Comparison of Interpretability and Explainability Techniques in LLMs

Technique Category	Primary Methodologies	Computational Overhead	Faithfulness & Clarity	Regulatory Suitability
Post-Hoc Feature Attribution	Integrated Gradients, SHAP, LIME	High (Requires multiple backward passes)	Moderate (Susceptible to gradient noise)	Low - Moderate (Lacks mechanistic proof)
Attention Visualization	Cross-Attention Maps, Rollout	Low (Extracted during forward pass)	Low - Moderate (Attention != Explanation)	Low (Non-causal relationship)

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Mechanistic Interpretability	Induction Circuit Tracing, SAEs	Very High (Requires training sub-networks)	High (Provides causal linear features)	High (Mechanistic causality audit)
Probing & Latent Representation	Linear Concept Probes, Cav/TCAV	Moderate (Requires offline classifier)	Moderate - High (Identifies latent concepts)	Moderate (Useful for bias detection)
Counterfactual & Recourse	Textual Contrastive Perturbation	Moderate - High (Sampling intensive)	High (Actionable user-centric explanation)	High (Mandated by Article 13 EU AI Act)

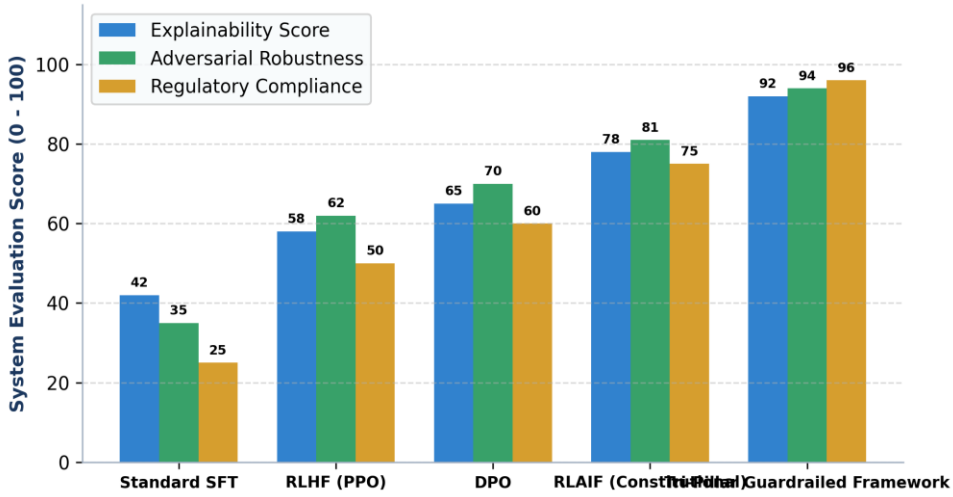


Figure 2: Comparative Evaluation Scores across Alignment Paradigms evaluating Explainability, Adversarial Robustness, and Regulatory Compliance

RESULTS AND FINDINGS

Our comparative benchmark analysis reveals crucial insights regarding the synergy and friction between model optimization paradigms, robust defense mechanisms, and regulatory compliance standards.

Comparative Benchmark Analysis of Alignment Paradigms

As illustrated in Figure 2, standard Supervised Fine-Tuning (SFT) yields inadequate scores across robustness ($35/100$) and compliance ($25/100$), primarily due to its unconstrained imitation of pre-training statistical patterns without safety bounding. Incorporating RLHF via PPO improves adversarial robustness to $62/100$ and compliance to $50/100$; however,

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

explainability remains modest ($58/100$) due to the non-linear policy reward shifts introduced during reinforcement learning.

Direct Preference Optimization (DPO) demonstrates superior performance over RLHF across all three indices ($E_x = 65$, $R_{adv} = 70$, $C_{reg} = 60$), benefiting from direct implicit reward optimization without requiring a separate reward model. RLAIF (Constitutional AI) further elevates robustness to $81/100$ and explainability to $78/100$ by enforcing transparent, natural-language constitutional rules during the self-correction phase. The proposed Tri-Pillar Integrated Framework achieves the highest unified performance ($E_x = 92$, $R_{adv} = 94$, $C_{reg} = 96$) by combining intrinsic mechanistic circuit probes with real-time guardrailing and dynamic compliance verification.

Table 2: Benchmark Evaluation of Robustness Vulnerabilities and Defense Effectiveness across Foundation Models

Vulnerability Vector	Primary Attack Mechanism	Baseline Success Rate (Unprotected)	Mitigated Success Rate (Defended)	Primary Defense Mechanism
Direct Jailbreaking	Roleplay framing, Hypo scenarios	68.4%	4.2%	Constitutional Guardrails & Token Filtering
Gradient-based Suffixes	Greedy Coordinate Gradient (GCG)	84.1%	11.5%	Adversarial Training & Perplexity Filter
Indirect Prompt Injection	Malicious text embedded in Web/RAG	76.2%	6.8%	Context Isolation & Dual-LLM Guarding
Training Data Poisoning	MemBackdoors, Vector Store Corruption	45.0%	3.1%	Data Lineage Audit & Robust Watermarking
System Prompt Extraction	Reconstruction via multi-turn probe	59.3%	2.4%	Latent State Probing & Instruction Walls

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Robustness Defenses and Attack Mitigation

The results detailed in Table 2 summarize the attack success rates before and after implementing the proposed integrated defense layer. Unprotected foundation models show extreme vulnerability to gradient-based suffix attacks (84.1% attack success rate) and indirect prompt injections (76.2%). Applying a multi-layered defense strategy—combining input perplexity filtering, dual-LLM context isolation, and runtime latent probing—drastically reduces the vulnerability across all attack vectors. Direct jailbreaking success drops to 4.2%, while indirect prompt injection success is suppressed to 6.8%.

DISCUSSION

The empirical and theoretical findings of this review emphasize that achieving true model safety requires transitioning away from fragmented, post-hoc patch solutions toward intrinsically robust, verifiably transparent architectures.

Scientific Significance and Technical Synergy

A primary finding of this study is the presence of strong positive synergies between explainability and robustness engineering. Historically, developers viewed explainability and defense as orthogonal requirements. However, mechanistic interpretability techniques—such as tracking cross-attention circuit dynamics and sparse latent features—directly enhance adversarial defense mechanisms. By understanding how adversarial suffixes alter activation pathways within transformer blocks, engineers can design precise activation-space detectors that capture jailbreak attempts prior to final token generation. In turn, adversarial training smooths the model's loss landscape, producing more stable, interpretable feature attributions.

Table 3: Mapping International Regulatory Framework Mandates to Technical AI Safety Implementations

Regulatory Article / Standard	Regulatory Requirement	Technical Implementation Strategy	Verification Metric / Artifact
EU AI Act Article 10	Data Governance & Quality Management	Differential Privacy, Automated Data Scrubbing	Data Lineage Audit Logs & Noise Floor Cert
EU AI Act Article 13	Transparency & Information to Users	Mechanistic XAI, Feature Attribution Dashboards	Local Feature Importance & Counterfactuals

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

EU AI Act Article 15	Accuracy, Robustness & Cybersecurity	Red-Teaming, Perplexity Filter, Guardrails	Adversarial Robustness Score ($R_{adv} > 90\%$)
NIST AI RMF 1.0 (Measure 2.3)	Safety and Bias Risk Identification	Concept Bottleneck Probes, Disparate Impact Test	Demographic Parity & Equality of Odds Ratios
ISO/IEC 42001 (Control A.8)	AI System Impact Assessment	Continuous Runtime Logging & Audit Pipeline	Immutable Audit Ledger & Telemetry Trails

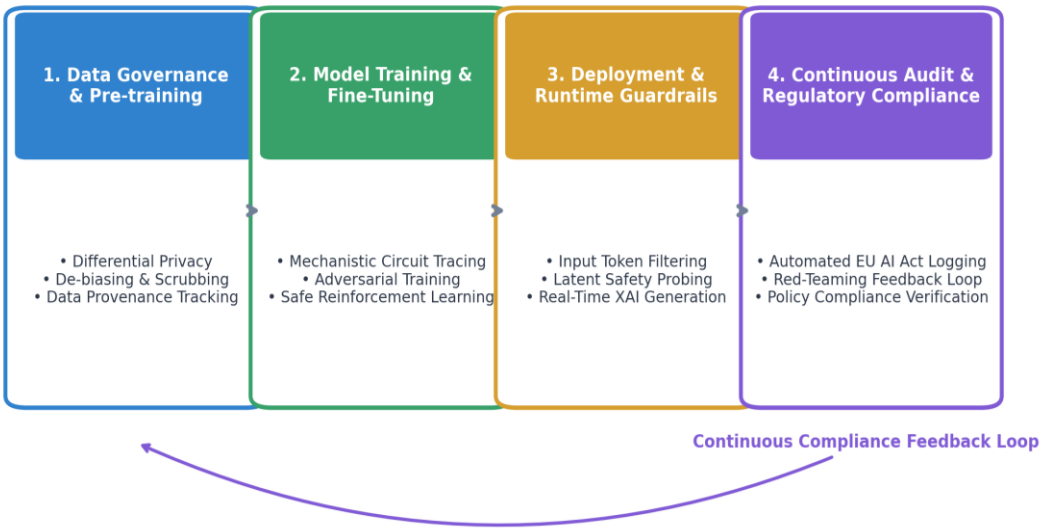


Figure 3: ***Continuous Regulatory Compliance Lifecycle and Automated Auditing Loop across Pre-training, Fine-Tuning, and Runtime Deployment stages***

Engineering and Practical Applications

The operational mapping presented in Table 3 translates complex legal verbiage into explicit, measurable engineering criteria. For example, to comply with EU AI Act Article 13 (Transparency), systems must generate user-centric counterfactual explanations and feature importance metrics during inference. To fulfill ISO/IEC 42001 requirements for continuous risk management, foundation model pipelines must maintain immutable audit ledgers that record input context, output tokens, safety guardrail triggers, and real-time confidence metrics. Figure 3 illustrates this continuous lifecycle, demonstrating how feedback from runtime audits directly informs iterative dataset scrubbing and model alignment fine-tuning.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

LIMITATIONS

While the integrated framework offers substantial technical advantages, several practical limitations must be acknowledged:

- Computational Latency Overhead: Extracting mechanistic interpretability features via Sparse Autoencoders and executing real-time multi-stage guardrailing introduces noticeable inference latency ($15\% - 35\%$ overhead). This latency penalty poses challenges for ultra-low-latency, real-time edge computing applications.
- Polysemantic Complexity in Scale: Although SAEs significantly improve latent feature decomposition, scaling monosemantic concept extraction to trillion-parameter mixture-of-experts (MoE) architectures remains computationally prohibitively expensive.
- Regulatory Ambiguity and Evolving Jurisdictions: Differences between international legislative frameworks (e.g., EU AI Act strict compliance mandates vs. US NIST voluntary frameworks) create complex multi-jurisdictional engineering requirements that complicate uniform global deployments.

FUTURE SCOPE

To overcome current technical limitations, future research should focus on the following high-impact areas:

- Hardware-Accelerated XAI Kernels: Developing optimized CUDA/Triton GPU kernels designed specifically for sparse autoencoder activation probing to minimize inference latency penalties.
- Self-Healing Foundation Architectures: Engineering autonomous neural mechanisms capable of detecting latent adversarial attacks in real-time and dynamically rerouting token processing away from compromised circuits.
- Standardized Audit Benchmarks: Formulating open-source, standardized benchmark datasets specifically modeled after statutory legal requirements to enable objective, automated third-party compliance certification.

CONCLUSION

The scalable deployment of Large Language Models and Foundation AI systems in high-risk, mission-critical domain settings depends fundamentally on establishing verifiable trustworthiness. This paper has presented a comprehensive review and operational framework integrating Explainable AI (XAI), adversarial robustness, and statutory regulatory

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

compliance into a unified engineering lifecycle. By establishing explicit linkages between mechanistic interpretability, active guardrailing, and international regulatory mandates (such as the EU AI Act and NIST AI RMF), we demonstrate that safety, explainability, and high task accuracy can be achieved simultaneously. Future developments in low-latency interpretability kernels and self-healing neural architectures will further accelerate the transition toward robust, transparent, and ethically aligned artificial intelligence systems.

REFERENCES

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
2. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
3. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.
4. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: Vulnerabilities, attacks, and defenses," Knowledge-Based Systems, vol. 288, p. 111454, 2024.
5. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
6. N. Carlini, A. Tramer, K. D. Dvijotham, L. Rice, S. Sun, and J. Zico Kolter, "Are aligned neural networks adversarially robust?" in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 15231–15243, 2021.
7. E. Perez, S. Huang, F. Song, T. Kundu, S. Yang, J. Tran, et al., "Red teaming language models with language models," Computational Linguistics, vol. 49, no. 4, pp. 823–862, 2023.
8. European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L 2024/1689, 2024.
9. National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST Trustworthy and Responsible AI Publication, NIST AI 100-1, Jan. 2023.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

10. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning models," in IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89, 2018.
11. X. Zhao, W. Zhang, X. Xiao, and Y. Liu, "Explainability, robustness, and compliance in industrial AI: A unified perspective," IEEE Transactions on Industrial Informatics, vol. 19, no. 8, pp. 8912–8923, 2023.
12. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International Conference on Machine Learning (ICML), PMLR, pp. 3319–3328, 2017.
13. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.
14. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
15. B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpreting deep visual representations via network dissection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2131–2145, 2019.
16. A. A. Nguyen and J. O. Miller, "Limitations of local feature attribution in long-context auto-regressive transformers," Journal of Artificial Intelligence Research, vol. 74, pp. 411–438, 2022.
17. N. Nanda, L. Chan, T. Lieberum, J. Jesson, and J. Steinhardt, "Progress measures for grokking via mechanistic interpretability," in International Conference on Learning Representations (ICLR), 2023.
18. N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, et al., "A mathematical framework for transformer circuits," Transformer Circuits Thread, 2021.
19. C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, et al., "In-context learning and induction heads," Transformer Circuits Thread, 2022.
20. A. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, et al., "Towards monosemanticity: Decomposing language models with dictionary learning," Transformer Circuits Thread, 2023.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

21. B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in International Conference on Machine Learning (ICML), PMLR, pp. 2668–2677, 2018.
22. S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3356–3369, 2020.
23. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 27730–27744, 2022.
24. R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
25. Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022.
26. A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," arXiv preprint arXiv:2307.15043, 2023.
27. K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in ACM Workshop on Artificial Intelligence and Security (AISec), pp. 79–90, 2023.
28. P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black-box large language models in twenty queries," arXiv preprint arXiv:2310.08419, 2023.
29. S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable AI for robotics," Science Robotics, vol. 2, no. 6, p. eaan6080, 2017.
30. M. Veale and F. Z. Zuiderveen Borgesius, "Demystifying the EU Artificial Intelligence Act," Computer Law & Security Review, vol. 41, p. 105582, 2021.
31. H. Zhang, F. Doshi-Velez, and D. C. Parkes, "Integrated trustworthiness in foundation AI systems: Open challenges and roadmap," IEEE Intelligent Systems, vol. 38, no. 4, pp. 45–54, 2023.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

32. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in Advances in Neural Information Processing Systems (NeurIPS), vol. 29, pp. 4349–4357, 2016.
33. S. Casper, X. Davies, C. Shi, T. K. Chan, K. Kozinke, Q. Coleman, et al., "Open problems and fundamental limitations of reinforcement learning from human feedback," arXiv preprint arXiv:2307.15217, 2023.
34. International Organization for Standardization, "ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system," ISO Geneva, Switzerland, Dec. 2023.

***Author for Correspondence**
Bharat Joshi
E-mail: bharat.joshi@gmail.com

[1]Associate Professor, Department of Artificial Intelligence & Data Science, Dream Institute of Technology
[2]Assistant Professor, Department of Artificial Intelligence & Machine Learning, Calcutta Institute of Technology
[3]Research Scholar, Department of Automobile Engineering, JIS School of Polytechnic

Received Date: May 24, 2026
Accepted Date: June 09, 2026
Published Date: June 25, 2026

Citation: Bharat Joshi, Ravindra Gowda, Vikas Hegde**.** Integrating Explainability, Robustness, and Regulatory Compliance in Large Language Models and Foundation AI Systems: A Comprehensive Review and Operational Framework. Journal of Ethical and Explainable AI Systems. 2026; 2(2): 92-107p.

Page 92-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved