International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

Journal Website: http://sciencebring.com/index.php/ijasr

Research Article

Adaptive Zero-Trust Authentication and Authorization for AI Agent Ecosystems Using SPIFFE/SPIRE

Submission Date: July 01, 2026, **Accepted Date:** August 15, 2026,
Published Date: September 14, 2026
Crossref doi: https/ doi.org/10.37547/ijasr-06-09-05

Rohan Malhotra
AI Research Scientist, India

Meera Iyer
Artificial Intelligence Research Analyst, India

Abstract

The emergence of autonomous and semi-autonomous AI agents is transforming distributed computing from conventional service-oriented architectures toward dynamic ecosystems in which software agents can independently invoke models, access data, communicate with other agents, and execute actions on behalf of users or organizations. This evolution creates a security problem that cannot be adequately addressed through static network boundaries or conventional identity mechanisms. Authentication and authorization must instead become continuous, workload-aware, and capable of distinguishing individual agent identities from transient execution contexts. This paper proposes an adaptive zero-trust authentication and authorization framework for AI agent ecosystems based on SPIFFE/SPIRE principles. The proposed approach conceptualizes every AI agent as an independently identifiable workload and combines cryptographically verifiable workload identity with contextual authorization, continuous trust evaluation, least-privilege policies, and adaptive risk decisions. The framework also incorporates computational-efficiency considerations because authentication, policy evaluation, telemetry, and model-mediated security controls can introduce additional processing overhead. The literature indicates that increasingly complex AI workloads require greater attention to computational efficiency, lifecycle management, transparency, and environmental cost (Bartoldson et al., 2023; Menghani, 2021; Schwartz et al., 2020). The study develops a conceptual architecture, identifies its functional components, and evaluates expected security and operational outcomes through analytical scenarios. Findings suggest that SPIFFE/SPIRE-oriented identity can provide a strong foundation for agent authentication, while adaptive authorization is necessary to address dynamic agent behavior, changing context, privilege escalation, and

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

inter-agent delegation. The framework contributes a structured security model for trustworthy agent ecosystems while identifying limitations related to policy complexity, identity lifecycle management, computational overhead, and the absence of standardized behavioral trust metrics.

KEYWORDS

Zero-Trust Security, AI Agents, SPIFFE, SPIRE, Workload Identity, Adaptive Authorization, Agentic AI, Identity Management, Risk-Based Access Control, Secure AI Infrastructure

INTRODUCTION

AI systems are increasingly moving beyond isolated prediction models toward distributed ecosystems composed of models, tools, APIs, data services, orchestration components, and autonomous agents. In such environments, an AI agent may retrieve information from one service, invoke another model, manipulate a database, communicate with another agent, or initiate an operational workflow. The resulting architecture differs fundamentally from traditional client-server applications because software components may dynamically assume roles, create execution chains, and make decisions with limited human intervention.

The security implications of this transformation are significant. Traditional authentication approaches frequently emphasize the identity of a human user, machine account, API key, or network location. Such assumptions become insufficient when multiple autonomous agents act simultaneously on behalf of users or organizations. A secure architecture must establish not only whether an agent possesses a valid identity but also whether that agent should be permitted to perform a particular operation under the current context.Zero-trust security provides a suitable theoretical foundation because it rejects implicit trust based on network position or previous authentication. Within an AI-agent environment, this principle can be operationalized by treating every agent workload as an independently identifiable entity whose access must be continuously evaluated. SPIFFE provides a workload-identity abstraction, while SPIRE can be conceptualized as the infrastructure responsible for issuing and managing workload identities. The application of this identity model to agentic systems has been specifically explored in the supplied work of Pappu, Bhushan, and Mittal, which establishes SPIFFE-based zero-trust authentication as a relevant direction for AI agent ecosystems (Pappu et al., 2025).

However, authentication alone does not establish comprehensive trust. An agent can possess a valid identity while still attempting an operation outside its intended role. Therefore, an effective architecture must combine strong workload authentication with contextual authorization, risk assessment, least privilege, continuous verification, and auditable policy enforcement.

A second challenge is operational efficiency. AI workloads can already impose substantial computational and environmental costs. Research on efficient deep learning, Green AI, carbon-aware computation, and AI lifecycle management

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

demonstrates that security mechanisms should not be designed independently of computational efficiency (Schwartz et al., 2020; Lannelongue et al., 2021; Verdecchia et al., 2023). Additional authentication, policy evaluation, monitoring, and cryptographic processing may create measurable overhead in high-volume agent ecosystems.

Accordingly, this research has four objectives: first, to formulate a conceptual SPIFFE/SPIRE-based identity architecture for AI agents; second, to extend workload authentication with adaptive authorization; third, to analyze the security and operational trade-offs of continuous trust evaluation; and fourth, to integrate computational-efficiency considerations into the proposed security model.

The scope of this paper is conceptual and analytical rather than experimental. It does not claim measured performance superiority over existing implementations. Instead, it establishes an architecture and evaluates its expected properties using the supplied literature and representative operational scenarios.

LITERATURE REVIEW

The literature supplied for this study can be divided into two principal bodies: research concerning AI computational efficiency and sustainability, and the more directly relevant work concerning SPIFFE-based zero-trust authentication for AI agents.

Schwartz et al. introduced the Green AI perspective by emphasizing that AI research should account for efficiency and environmental cost rather than focusing exclusively on predictive performance. Their work provides an important theoretical basis for this paper because security architecture should similarly consider resource consumption as an architectural quality attribute (Schwartz et al., 2020). Strubell et al. highlighted the energy and policy implications associated with deep learning workloads, demonstrating that computational intensity can have consequences extending beyond technical performance (Strubell et al., 2019). Lacoste et al. further demonstrated the importance of quantifying carbon emissions associated with machine-learning computation, reinforcing the argument that additional infrastructure functions should be evaluated according to their computational implications (Lacoste et al., 2019).

Research on computational efficiency provides another relevant perspective. Bartoldson et al. examined algorithmic approaches for reducing the computational requirements of deep learning, while Menghani surveyed techniques for making deep-learning models smaller, faster, and more efficient (Bartoldson et al., 2023; Menghani, 2021). Canziani et al. analyzed deep neural network models in practical applications, illustrating the importance of understanding model-level efficiency when AI systems are deployed operationally (Canziani et al., 2016). Li's work on training large models followed by compression further illustrates the broader architectural principle that computational resources should be allocated strategically across AI lifecycles (Li, 2020).

The lifecycle dimension is particularly important for agent ecosystems. Haakman et al. argued that conventional AI lifecycle models require revision in response to contemporary operational conditions. This insight can be extended to security: if AI

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

lifecycle models become dynamic, then identity and authorization mechanisms must also accommodate dynamic creation, deployment, updating, and retirement of AI workloads (Haakman et al., 2021).

Dodge et al. emphasized improved reporting of experimental results, an observation that has methodological implications for security research. Claims about an adaptive authorization architecture should be supported by clearly defined evaluation criteria rather than broad assertions of security improvement (Dodge et al., 2019). Dodge's research on the carbon intensity of AI in cloud environments similarly indicates that infrastructure-level decisions have measurable computational consequences (Dodge, 2022).

Patterson examined the future trajectory of machine-learning training carbon footprints, while Gupta investigated the environmental footprint of computing. Together, these studies reinforce the proposition that large-scale AI infrastructure must balance functionality with resource efficiency (Patterson, 2022; Gupta, 2023). Xu et al. and Verdecchia et al. further demonstrate the growing importance of systematic approaches to Green AI and efficient AI development (Xu et al., 2021; Verdecchia et al., 2023).

The directly relevant security foundation is provided by Pappu, Bhushan, and Mittal, who investigated SPIFFE-based zero-trust authentication for AI agent ecosystems (Pappu et al., 2025). Their work establishes workload identity as an important mechanism for moving AI-agent authentication beyond conventional user-centric approaches. Building on this foundation, the present paper focuses specifically on the additional problem of adaptive authorization: determining what an authenticated agent should be allowed to do under changing contextual and risk conditions.

The principal research gap therefore lies at the intersection of these areas. Existing supplied literature provides substantial foundations for computational efficiency, sustainability, AI lifecycle management, and SPIFFE-oriented authentication, but it does not provide a complete integrated model in which workload identity, contextual authorization, adaptive risk evaluation, and computational efficiency are treated as interconnected architectural concerns. This paper addresses that gap through a conceptual framework.

METHODOLOGY

3.1 Research Design

This study uses a conceptual architecture and analytical research methodology. Rather than implementing a production SPIRE deployment or conducting a benchmark experiment, the research derives architectural principles from the supplied literature and translates them into an adaptive security model for AI agent ecosystems.

The methodology follows four analytical stages: identification of security requirements, construction of a workload-identity model, development of an adaptive authorization mechanism, and evaluation of expected security and operational outcomes.

The approach is particularly appropriate because AI agent ecosystems remain heterogeneous. Agents may differ in purpose, computational environment, privileges, data sensitivity, and autonomy. A single

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

static authorization mechanism would therefore be unlikely to accommodate all deployment scenarios.

3.2 SPIFFE/SPIRE-Based Workload Identity Model

The proposed architecture begins with workload identity. Each AI agent is represented as an independently identifiable workload rather than being treated merely as an extension of a human user.

Conceptually, the identity lifecycle contains four stages: identity registration, credential issuance, identity verification, and credential renewal or revocation. SPIFFE provides the conceptual identity abstraction, while SPIRE provides the identity-management infrastructure.

For example, consider a research organization operating three AI agents: a retrieval agent, an analysis agent, and a reporting agent. The retrieval agent may access approved information repositories, the analysis agent may process retrieved information, and the reporting agent may write results to a document system. Although all three agents may operate within the same organizational network, each should possess an independently verifiable workload identity.

This separation limits the assumption that location implies trust. A compromised retrieval agent should not automatically inherit the permissions of the analysis or reporting agent.

Pappu et al. demonstrate the relevance of SPIFFE-based identity to zero-trust authentication in AI agent ecosystems (Pappu et al., 2025). The present framework extends that principle by separating authentication from authorization. Identity establishes who or what the workload is; authorization determines what that workload is allowed to do.

3.3 Continuous Authentication

The second component is continuous authentication. Traditional authentication may occur when an agent initially connects to a service. In an autonomous ecosystem, however, an agent can remain active for extended periods and may interact with multiple services.

The proposed model therefore evaluates identity at significant trust boundaries. Verification can occur when an agent:

1. initiates a new service interaction;
2. requests a privileged operation;
3. accesses a higher-sensitivity dataset;
4. delegates work to another agent; or
5. exhibits a changed execution context.

The objective is not to repeatedly perform expensive operations without reason, but to create verification points proportional to risk.

This distinction is important for computational efficiency. Excessive security checks may increase latency and resource consumption, while insufficient verification may create security exposure. Green AI research supports the broader principle that computational costs should be explicitly considered rather than treated as irrelevant infrastructure overhead (Schwartz et al., 2020; Lannelongue et al., 2021).

3.4 Adaptive Authorization Model

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

Authentication is followed by authorization. The proposed authorization decision can be represented conceptually as:

Authorization Decision = f(Identity, Resource, Action, Context, Risk, Trust State)

Identity identifies the agent. Resource identifies the requested object or service. Action specifies the operation. Context incorporates factors such as execution environment, request origin, task state, and delegation chain. Risk represents the estimated potential impact of granting the request.

The model supports four principal decision outcomes:

Allow: The requested operation satisfies identity, policy, and contextual requirements.

Allow with restrictions: The operation is permitted but constrained, such as read-only access or restricted data fields.

Step-up verification: Additional evidence is required before execution.

Deny: The operation violates policy or exceeds the acceptable risk threshold.

This model is preferable to identity-only authorization because a legitimate agent can become risky under a changed context. For instance, an analysis agent may normally access anonymized research data but should not automatically receive permission to export raw personal records merely because it possesses a valid identity.

3.5 Risk-Adaptive Trust Evaluation

The proposed framework treats trust as dynamic rather than permanent. A conceptual risk score can be expressed as:

$$R = w_1I + w_2A + w_3C + w_4D + w_5B$$

where:

- I represents identity confidence;
- A represents requested action sensitivity;
- C represents contextual risk;
- D represents delegation-chain risk; and
- B represents behavioral anomaly indicators.

The weights are deployment-specific and should be empirically calibrated in future research.

An agent requesting a low-risk read operation may receive a low risk score and be authorized automatically. The same agent requesting administrative privileges or attempting to access a sensitive resource may generate a substantially higher score.

This design creates a direct connection between zero-trust principles and AI autonomy: authorization becomes a decision about the current operation rather than a permanent property of the agent.

3.6 Least-Privilege and Delegation Control

Agent ecosystems introduce a distinctive challenge: delegation. An agent may invoke another agent to complete a task. If privileges are transferred without restriction, the resulting delegation chain can unintentionally amplify authority.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

The proposed architecture therefore applies privilege attenuation. A delegated agent should receive only the minimum privileges required for the delegated operation.

For example, if Agent A can retrieve customer analytics but delegates summarization to Agent B, Agent B should receive access only to the necessary analytical representation rather than the complete underlying customer database.

The delegation chain should remain cryptographically and logically traceable. This provides an auditable relationship between the original request, intermediate agents, and final action.

3.7 Policy Enforcement Architecture

The framework consists conceptually of five interacting layers.

The Identity Layer establishes workload identities.

The Authentication Layer verifies that an agent possesses a valid identity and corresponding credentials.

The Policy Layer determines whether a requested action complies with organizational rules.

The Risk Layer evaluates contextual and behavioral conditions.

The Enforcement and Audit Layer applies the final decision and records security-relevant events.

This separation reduces coupling between identity management and business authorization. It also allows policy logic to evolve without requiring fundamental changes to workload identity.

3.8 Computational-Efficiency Considerations

Security controls create operational costs. Certificate or credential validation, policy evaluation, telemetry generation, risk calculation, and audit logging can increase CPU, memory, network, and storage requirements.

The framework therefore proposes risk-proportional security. Low-risk interactions should use efficient validation paths, whereas high-risk operations can trigger deeper evaluation.

This principle is consistent with the literature on efficient AI, which emphasizes that performance and resource consumption must be considered together (Bartoldson et al., 2023; Menghani, 2021). Similarly, Green AI research demonstrates that computational cost is an important architectural consideration (Verdecchia et al., 2023).

An efficient implementation should therefore minimize redundant policy evaluations, cache non-sensitive policy decisions where appropriate, use short but practical verification paths, and avoid generating unnecessary telemetry.

3.9 Analytical Evaluation Criteria

The proposed framework is evaluated against six conceptual criteria: authentication strength, authorization precision, least-privilege enforcement, adaptability, auditability, and computational efficiency.

The methodology does not assign experimentally measured numerical values because no implementation or benchmark dataset was supplied. Instead, the evaluation determines whether the architecture logically satisfies each criterion and identifies potential trade-offs.

RESULTS

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

ISSN-2750-1396

The conceptual evaluation indicates that the proposed SPIFFE/SPIRE-oriented architecture can address several weaknesses associated with conventional identity-centric security models for AI agents. The first finding is that workload-level identity provides a more appropriate authentication abstraction for autonomous agents than relying exclusively on human or application-level identities. Each agent can be treated as an independently verifiable workload, allowing security policies to distinguish between agents even when they operate within the same infrastructure. This finding is consistent with the supplied SPIFFE-based AI-agent security research, which positions workload identity as a foundation for zero-trust authentication (Pappu et al., 2025).

The second finding concerns the separation between authentication and authorization. A valid workload identity does not necessarily justify every requested operation. By evaluating resource sensitivity, requested action, execution context, delegation relationships, and risk, the proposed framework provides finer-grained control than identity-only access. This is particularly important for autonomous systems because an agent may remain legitimate while its requested operation becomes inappropriate.

Third, adaptive authorization improves resilience against privilege escalation. In the proposed model, authorization is recalculated when risk-relevant conditions change. An agent that is authorized to read information during normal operation can be restricted when it attempts an unusual administrative action. The model therefore reduces the possibility that a compromised or misdirected agent can exploit previously granted privileges indefinitely.

Fourth, delegation-aware authorization emerges as a critical requirement. Agent-to-agent communication creates a potential privilege-propagation problem. The framework addresses this through privilege attenuation, requiring delegated agents to receive only those permissions necessary for their assigned task. This produces a more constrained trust chain and improves auditability.

Fifth, the analysis identifies computational overhead as a central trade-off. Continuous verification and contextual policy evaluation can strengthen security but may increase latency and resource utilization. This issue is especially relevant for AI systems because AI computation itself can be resource intensive. Research on computational efficiency and Green AI indicates that resource utilization should be treated as an important system-level concern rather than an incidental implementation detail (Schwartz et al., 2020; Lannelongue et al., 2021). The proposed risk-proportional verification mechanism consequently provides a mechanism for balancing security depth against operational cost.

Sixth, the framework improves auditability by connecting agent identity, authorization decision, resource, action, and delegation context. Such traceability is valuable for incident investigation and governance because an organization can reconstruct which workload performed an operation and under which authorization conditions.

The results also reveal important limitations. The framework depends on accurate workload identification and reliable policy configuration. Incorrect policies can deny legitimate actions or permit dangerous operations. Similarly, behavioral

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

risk indicators may generate false positives or false negatives if they are poorly calibrated. The model therefore provides an architectural foundation rather than a guarantee of security.

Overall, the findings suggest that SPIFFE/SPIRE-based workload identity is most effective when used as the authentication foundation of a broader adaptive authorization system. Authentication establishes verifiable workload identity, while contextual authorization, risk assessment, delegation control, and continuous policy enforcement transform identity into practical zero-trust access control.

Discussion

The central theoretical contribution of the proposed framework is the integration of workload identity with adaptive authorization. Conventional authentication models often treat successful identity verification as the principal security event. For autonomous AI ecosystems, this assumption is inadequate because the same agent can execute operations with substantially different security consequences. The framework therefore conceptualizes trust as a contextual and continuously evaluated property rather than a static characteristic.

This approach extends the SPIFFE-based zero-trust direction identified by Pappu et al. by introducing authorization as an independent decision layer (Pappu et al., 2025). The distinction is significant. SPIFFE/SPIRE-oriented identity can establish that an agent corresponds to an authorized workload, but organizational security requirements also depend on what the workload is attempting to accomplish. Adaptive authorization consequently becomes the mechanism that translates verified identity into bounded operational authority.

The framework also has implications for AI lifecycle management. Haakman et al. argued that AI lifecycle models require revision as AI systems become more complex and operationally integrated (Haakman et al., 2021). An equivalent security lifecycle should account for agent registration, identity issuance, policy assignment, privilege modification, delegation, monitoring, suspension, and retirement. An agent that is safe at deployment may become inappropriate after a software update, task change, or alteration in its operating environment. Identity management must therefore be lifecycle-aware.

A major trade-off concerns security versus computational efficiency. Stronger security generally requires more verification, policy processing, monitoring, and auditing. Yet AI workloads already have substantial computational demands. Research into efficient deep learning and Green AI establishes that resource consumption is an important dimension of AI system quality (Bartoldson et al., 2023; Menghani, 2021; Verdecchia et al., 2023). A security framework that introduces excessive computational overhead could therefore create operational inefficiency even while improving security.

The proposed risk-proportional approach addresses this contradiction by allocating security effort according to operation sensitivity. Low-risk operations can follow efficient verification paths, while high-risk actions can trigger more comprehensive evaluation. This does not eliminate overhead, but it provides a rational mechanism for controlling it.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

ISSN-2750-1396

Another important issue is policy complexity. Adaptive authorization introduces additional decision variables, including context, delegation, resource sensitivity, and behavioral signals. As the number of agents and resources increases, policy interactions may become difficult to reason about. Policy conflicts can create either excessive restriction or unintended privilege. Consequently, future implementations require policy validation, automated testing, and explainable authorization decisions.

The framework also has limitations concerning behavioral trust. An agent's unusual behavior does not necessarily imply malicious behavior. Autonomous systems may legitimately perform unexpected operations because of changing tasks or environmental conditions. Conversely, a compromised agent may behave within expected patterns while executing a harmful objective. Behavioral indicators should therefore complement, rather than replace, strong identity and policy controls.

The sustainability dimension further broadens the discussion. Research on carbon emissions, computational intensity, and Green AI suggests that security infrastructure should be evaluated as part of the overall AI computing lifecycle (Lacoste et al., 2019; Patterson, 2022). Security telemetry, cryptographic operations, policy engines, and monitoring services consume resources. Consequently, future research should measure the security benefit obtained per unit of computational and energy overhead.

Finally, the framework highlights a governance requirement: authorization decisions should be explainable to system operators. When an agent is denied access, organizations should be able to determine whether the decision resulted from identity status, policy violation, elevated risk, delegation constraints, or contextual conditions. Explainability is therefore not merely a usability feature; it is essential for operational accountability.

CONCLUSION

This paper proposed an adaptive zero-trust authentication and authorization framework for AI agent ecosystems using SPIFFE/SPIRE-oriented workload identity. The central argument is that autonomous agents require security mechanisms capable of distinguishing workload identity from operational authority. A valid identity establishes authenticity but does not independently determine whether a particular action should be permitted.

The proposed architecture addresses this distinction through continuous authentication, contextual authorization, risk-adaptive decisions, least-privilege enforcement, delegation control, policy enforcement, and security auditing. SPIFFE/SPIRE provides the conceptual foundation for workload identity, while adaptive authorization extends that foundation toward dynamic access control for autonomous agent environments. The supplied SPIFFE-focused research provides direct support for the relevance of this direction (Pappu et al., 2025).

The research also identifies computational efficiency as an important secondary design requirement. AI systems already involve substantial computational workloads, and additional authentication and authorization processes can increase infrastructure cost. The Green AI and efficient-computing literature therefore supports incorporating resource

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

efficiency into the design of security architectures rather than treating it as an independent concern (Schwartz et al., 2020; Verdecchia et al., 2023).

The principal contribution is a conceptual integration of identity, authorization, context, risk, delegation, and efficiency into a unified security model. Its limitations include the absence of empirical benchmark results, standardized behavioral trust metrics, and large-scale implementation validation. These limitations define several opportunities for future research.

Future studies should implement the framework using real SPIFFE/SPIRE deployments and evaluate authentication latency, authorization throughput, policy evaluation overhead, credential-renewal performance, and resource consumption. Additional research should investigate machine-readable authorization policies, agent delegation graphs, behavioral anomaly detection, policy-conflict analysis, and energy-aware security controls. Empirical evaluation should also follow transparent reporting practices so that security improvements can be evaluated alongside computational and operational costs, consistent with the broader methodological concerns raised in the supplied literature (Dodge et al., 2019).

Ultimately, secure agentic AI requires more than authenticating autonomous workloads. It requires continuously determining whether an authenticated workload should be trusted to perform a specific action, under a specific context, against a specific resource, with a specific level of authority. A SPIFFE/SPIRE-based adaptive zero-trust architecture provides a promising conceptual foundation for achieving this objective while preserving the principles of least privilege, continuous verification, and responsible computational use.

REFERENCES

1. D. Amodei and D. Hernandez, “AI and compute,” Accessed: Aug. 3, 2023. [Online]. Available: https://openai.com/blog/ai-and-compute/
2. B. R. Bartoldson, B. Kailkhura, and D. Blalock, “Compute-efficient deep learning: Algorithmic trends and opportunities,” J. Mach. Learn. Res., vol. 24, pp. 1–77, 2023.
3. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” in Proc. ACM Conf. Fairness, Accountability, Transparency, Virtual Event, 2021, pp. 610–623.
4. A. Canziani, A. Paszke, and E. Culurciello, “An analysis of deep neural network models for practical applications,” 2016, arXiv:1605.07678.
5. J. Dodge, “Measuring the carbon intensity of AI in cloud instances,” in Proc. ACM Conf. Fairness, Accountability, Transparency, Seoul, South Korea, 2022, pp. 1877–1894.
6. J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, “Show your work: Improved reporting of experimental results,” 2019, arXiv:1909.03004.
7. J. Gitzel, M. Platenius-Mohr, and A. Burger, “Estimating the sustainability of AI models based on theoretical models and experimental data,” 2023, arXiv:202301.0406.v1.
8. U. Gupta, “Chasing carbon: The elusive environmental footprint of computing,” in Proc. IEEE Int. Symp. High- Perform. Comput. Archit., Montreal, QC, Canada, 2023, pp. 854–867.

International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 09 Pages:51-62
OCLC – 1368736135

9. M. Haakman, L. Cruz, H. Huijgens, and A. van Deursen, "AI lifecycle models need to be revised: An exploratory study in fintech," Empirical Softw. Eng., vol. 26, pp. 1–29, Jul. 2021, doi: 10.1007/s10664-021-09993-1.
10. D. Hershcovich, N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold, "Towards climate awareness in NLP research," 2022, arXiv:2205.05071.
11. L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," Adv. Sci., vol. 8, no. 12, pp. 1–10, 2021.
12. A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," 2019, arXiv:1910.09700.
13. Z. Li, "Train big, then compress: Rethinking model size for efficient training and inference of transformers," in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 5958–5968.
14. G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," ACM Comput. Surv., vol. 55, no. 12, pp. 1–37, 2021, doi: 10.1145/3578938.
15. D. Patterson, "The carbon footprint of machine learning training will plateau, then shrink," Computer, vol. 55, no. 7, pp. 18–28, 2022, doi: 10.1109/MC.2022.3148714.
16. K. Pappu, B. Bhushan and A. Mittal, "SPIFFE-Based Zero-Trust Authentication for AI Agent Ecosystems," 2025 International Conference on Computer and Applications (ICCA), Bahrain, Bahrain, 2025, pp. 1-7, doi: 10.1109/ICCA66035.2025.11431026.
17. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Commun. ACM, vol. 63, no. 12, pp. 54–63, Nov. 2020, doi: 10.1145/3381831.
18. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," 2019, arXiv:1906.02243.
19. R. Verdecchia, R. Sallou, and L. Cruz, "A systematic review of Green AI," Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery, Wiley Online Library, 2023, Art. no. e1507.
20. R. Verdecchia, L. Cruz, J. Sallou, M. Lin, J. Wickenden, and E. Hotellier, "Data-centric Green AI an exploratory empirical study," in Proc. Int. Conf. ICT Sustainability, Plovdiv, Bulgaria, 2022, pp. 35–45.
21. J. Xu, W. Zhou, Z. Fu, H. Zhou, and L. Li, "A survey on green deep learning," 2021, arXiv:2111.05193.
22. D. Rydning, J. Reinsel, and J. Gantz, "The digitization of the world from edge to core," Framingham: Int. Data Corporation, vol. 16, pp. 1–28, 2018.