ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

ARIA: ADAPTIVE RECTIFIED INTEGRATED ATTENTION — A THEORETICALLY GROUNDED MULTI-STREAM TRANSFORMER ARCHITECTURE RESOLVING SYSTEMIC INSTABILITIES FOR SECURE COMMON INTENT ORCHESTRATION IN SCALABLE LANGUAGE MODELING

*** Lakshya Singh, **Akriti Vishwakarma, ***Arpita Yadav & ****Aditya Naikwadi***

, ** & * B.Sc Information Technology Student, Department of Computer Science and Information Technology*
***** M.Sc. Data Science and Big Data Analytics Student, Department of Computer Science & Information Technology, B.K. Birla College, (Empowered Autonomous Status), Kalyan.*

Abstract:

The ARIA transformer architecture is revealed as a viable solution to the Common Intent Orchestration (CIO) framework, addressing the stability and causality issues of multi-agent systems. The use of multi-stream sigmoid attention models, while enabling long-context tasks, is often marred by problems such as gradient instability and autoregressive violations. The ARIA model overcomes all 18 critical failures of such approaches through a set of six strategic innovations, including L2 Normalised Sigmoid Attention, Content Relative RoPE, and Lagged State Memory. The use of Entropy Variance Monitors for parallel repair and Precision Anchor Tokens ensures structural stability without incurring computational bloat. Complexity analysis of the model indicates a highly efficient time complexity of O(N^1.2) and a moderate 1.3x KV cache overhead. The ARIA model meets all ten architectural requirements for a reliable agentic orchestration, a major leap in the field of high-stability large-scale language modeling.

Introduction:

As the field of Artificial Intelligence continues to shift towards secure multi-agent orchestration, the Common Intent Orchestration (CIO) framework emerged as a critical framework to ensure semantic consistency in multi-hop reasoning processes. However, current production deployments of the framework face a significant bottleneck in the structural limitations of sigmoid-based multi- stream architectures. These architectures, while meeting the subquadratic complexity requirement for long-context tasks, fail to meet the output stability and causal validity requirements for agentic pipelines.

In an attempt to overcome the limitations of current sigmoid-based architectures, we propose a novel framework, ARIA, which represents a significant architectural shift in the field of Artificial Intelligence, meeting the output stability requirement while maintaining cycle-free gradients and strict causal validity.

Key Technical Contributions:

- Architectural Stability: A novel output-bounded guarantee for sigmoid-based attention mechanisms.
- Causal Integrity: Strict adherence to autoregressive generation, a critical requirement that was previously unmet in multi-stream architectures.

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

- CIO Enablement: The proposed framework meets all ten architectural requirements for production-grade Common Intent Orchestration.

Statement of Problem:

The current implementation of multi-stream sigmoid attention is fundamentally incompatible with the Common Intent Orchestration (CIO) architecture due to 18 systemic failure modes. These include the catastrophic sigmoid output explosion, in which outputs grow at a rate of O(N), leading to a 9,000 times numerical explosion in 10,000 token contexts. These instabilities combine with the positional encoding incompatibilities to threaten the long-range reasoning and intent verification that multi-agent governance requires.Collectively, these architectures also suffer from causal leaks and cyclic gradient dependencies that threaten model integrity. For example, during hierarchical context compression, "future" tokens are accessed during computation, leading to silent corruption that cannot be captured by standard training metrics. These architectural failures combine to break reverse-mode autodiff and violate autoregressive validity in a manner that is fundamentally unmathematical and unfit for agentic pipelines.

Significance of study:

This research creates a vital link between theoretical stability and applied AI governance, tackling the "building on sand" problem wherein architectural issues compromise the effectiveness of higher-level security. The key innovation of this research is the development of a comprehensive taxonomy of 18 failure modes, which creates a framework for future LLM audits. From a mathematical perspective, the output boundedness of the model is formally proven through the NSIG theorem, which demonstrates sigmoid attention stability, a previously unheard-of achievement in the field. In terms of hardware, this research makes a major efficiency gain, cutting KV cache overhead from 3.0x to 1.3x, which corresponds to a memory savings of 83 GB in standard A100 clusters for 128K token sizes. In terms of software, this research creates a standardized four-stage training pipeline, which streamlines complex hyperparameter management. Ultimately, this research makes the sigmoid-based multi-stream transformer, previously a highly unstable and experimental model, a production-ready platform for reliable multi-agent orchestration.

Limitations of the study

Despite the strong theoretical underpinnings of ARIA, the current status of the model as an analytical proposal creates a number of empirical risks, beginning with the fact that the model currently lacks large-scale benchmarking, as the \$O(N^{1.2})\$ complexity and KV cache efficiencies were determined through formal specification and may be affected by real-world hardware constraints such as memory bandwidth. In addition, the mechanism of the Causal Pyramid Compression algorithm creates a latency overhead that should be further evaluated to optimize the relative costs of the algorithm.

The Lightweight Entity Tagging model, based on the BiLSTM-CRF architecture, is currently susceptible to performance degradation in extreme domain shifts. In addition, the theoretical 20- 30-30-20% compute budget for the four-stage training curriculum has yet to be optimized for different model sizes and hardware configurations. Finally, the relative interaction of the training protocol and the data curriculum itself is an open

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

question, marking the transition from formal proof to production-ready implementation as a key area of future research.

Objectivity of study:

The major goals and objectives of this research have been stated as follows:

i. To perform a systematic audit of the failure modes in multi-stream sigmoid attention architectures and propose a rigorous taxonomy that differentiates between Class A structural failures and Class B optimization failures.

ii. To propose and formally define six architectural components: CR-RoPE, NSIG, CPC with CST and PAT, LET, EVM-PR, and LSM that cover all eighteen failure modes.

iii. To formally prove output boundedness (NSIG), causal validity (CST), positional encoding coherence (CR-RoPE), and reasoning detection (EVM) using formal mathematical methodologies.

iv. To rigorously prove that ARIA has subquadratic time complexity: O(N^1.2) in typical scenarios and O(N^1.5) in worst-case scenarios, and that KV cache costs are reduced to 1.3 times the baseline.

v. To substantiate that ARIA is the first multi-stream sigmoid architecture that satisfies all ten requirements for production-grade scalable language models, thereby meeting the architectural prerequisites for CIO-based multi-agent governance.

Hypothesis of the study:

H1 (Structural Integration): ARIA's use of L2-normalized attention and causally valid compression results in a sub-quadratic time complexity of O(N^1.2), which ensures the stability of the output and the autoregressive integrity necessary for effective multi-agent orchestration.

H2 (Architectural Redesign): The paper draws a clear distinction between Class A (structural) and Class B (optimization) inefficiencies, arguing that the earlier mistakes in multi-stream design were fatal and that a complete architectural redesign is necessary, rather than attempting to patch up the problem through training.

H3 (Positional Coherence): The addition of Content Relative RoPE helps to overcome the dual gradient conflicts, thereby greatly improving the model's ability to track long-range dependencies in complex, multi-hop reasoning tasks.

H4 (Memory Efficiency): The use of shared value projections in the attention streams reduces the KV cache cost from 3.0x to 1.3x, allowing for 128K token context inference on distributed GPU hardware.

Review of literature:

The development of transformer architectures has progressed from the basic scaled dot-product attention mechanism (Vaswani et al., 2017) towards the development of specific mechanisms for long-context efficiency. Although Softmax normalization was initially effective for numerical stability, it also led to "attention dilution" for long sequences. FlashAttention (Dao et al., 2022) was later introduced for memory pattern optimization, resulting in $O(N)$ complexity, while sparse approaches like Longformer have also achieved sub-quadratic complexity, but all of these have faced the limitations of the Softmax approach. Recent research has focused more on the use of sigmoid-based approaches for the development of the attention mechanism, removing the "attention dilution" issue, as described by Ramapuram et al. (2024). Some of the recent key developments for the numerical stability of the basic attention mechanism have been the fixed bias stabilization, where \$b = -\log(n)\$, introduced by Sun et al. (2023), and the query-key normalization, introduced by Henry et al. (2020), which deals with the early stages of gradient variance. Hybrid compression approaches like PHOTON have also improved the "bottom-up" memory approach, making way for the development of the ARIA, which combines all the recent developments of the basic attention mechanism.

The development of positional encodings has moved away from absolute forms to relative approaches such as RoPE (Su et al., 2024), which utilizes vector rotations for length generalization, and ALiBi (Press et al., 2022), which utilizes additive bias for parameter-free extrapolation. These fundamental techniques form the basis for the development of the current techniques for the verification of reasoning, with Chain of Thought and Iterative Reflection (Shinn et al., 2023) improving the performance of multi-hop tasks through fine-grained self-evaluation.

To ensure the stability of the training process, the recent developments incorporate the use of entropy-based attention analysis to detect any error in the process of reasoning, a critical component of the ARIA Entropy Variance Monitor. However, despite the developments, the use of the common intent orchestration (CIO) framework (Krishnan & Mehta, 2025), which criticizes the use of the standard sigmoid-based approach for lacking the critical architectural security that the ARIA model provides, indicates the importance of the structural changes introduced.

Research Methodology:

The research is conducted through a theory-driven methodology that consists of two critical phases, with a focus on structural integrity rather than empirical measures.

Phase 1 of the methodology is called the Systematic Failure Audit, which is a systematic evaluation of previous multi-stream architectures in terms of autoregressive correctness and numerical stability. This phase categorizes failure into Class A, which includes Structural issues such as causality and boundedness, and Class B, which refers to issues related to Optimization. This categorization includes a total of 18 failure modes, which are related to attention and memory.

Phase 2 of the methodology is called Architectural Specification, which is a focus on the implementation of mathematical interventions to rectify these Class A inconsistencies. This is done while ensuring that the modifications are filtered to eliminate specific failure modes without creating secondary gradient and causality conflicts. The resulting model is called the ARIA model, which is a highly precise specification that ensures attention normalization, positional encoding, and memory are harmonized from a mathematical perspective to meet the stringent requirements of the CIO framework.

The last part of the research methodology is Formal Complexity Analysis, which goes beyond the asymptotic theory and enters the domain of concrete operational estimates. In this context, the research confirms that the dominant terms of ARIA have sub-quadratic complexity, and the cache requirements of KV are highly optimized in terms of projection sizes. This is a rigorous approach to ensuring that the architecture is efficient and effective,

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

particularly in terms of the high-density requirements of the CIO framework.

In order to move from theory to practice, a four-stage training pipeline is used. This is a training protocol that prioritizes signal stability, ensuring that bounded mechanisms are introduced after normalization. This is a rigorous approach to addressing the problem of signal stability, as the researchers were able to reduce the number of hyperparameters from more than ten to four, ensuring that the parameters are locked to the architectural scaling constraints. This is a recognition that, while mathematical integrity is guaranteed, the actual operational parameters require empirical fine-tuning to accommodate the nuances of long context deployment.

Data Analysis and Interpretation:

The analytical framework provided by ARIA's design clearly showcases its superiority over existing models through four critical domains. Firstly, the Failure Taxonomy categorizes six Class A structural failures, including sigmoid output explosion (S_3) and causality violations. This validates the argument that existing multi-stream models require a complete redesign rather than tuning. Complexity Analysis validates ARIA's three-stream model achieving a typical $O(N^{1.2})$ complexity. Furthermore, the shared value projection significantly reduces KV cache costs from 192 GB to 83 GB at 128K tokens and can thus be deployed on existing A100 clusters. Thus, ARIA is the first model to achieve all ten production-grade requirements and surpasses both softmax and existing sigmoid models. Its four-stage training protocol (20/30/30/20%) methodically adds components from stable init to full context extension. This guarantees the presence of bounded signals (NSIG) before enabling reasoning repair (EVM). This methodical evolution, combined with the reduction to a four-hyperparameter set, provides a mathematically sound and operationally feasible platform for the CIO.

Challenges:

There were also a number of significant challenges that were faced during the course of this research.

First Challenge:

First and foremost, there was the challenge in the systematic identification and characterization of architectural failure modes before the start of the design process. Most studies focused on one or two failure modes separately. For this research, therefore, the challenge was in the identification of the interactions. For example, the identification that solving the NSIG output explosion (S3) is a prerequisite to solving the ECM monitoring stability (O8) because the latter is only possible for a bounded signal was a challenge.

Second Challenge:

The second challenge was in the design of the CST and PAT architectures to facilitate causally valid hierarchical compression without loss in compression utility. While the top-down summarization passes used in the previous architectures were very efficient in that they allow for the utilization of the whole document in the creation of the chunk summary, the bottom-up approach had to sacrifice this for efficiency. While the solution adopted in ARIA using CST as a causal prefix to the next chunk and bypassing the compression for precision-critical tokens using PAT is a pragmatic solution, it is to the expense of 5-15% more tokens.

Third Challenge:

The third challenge is the cycle elimination in the LSM. While it is naturally desirable to bridge contexts between

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

layers in multi-hop reasoning, any model that allows layer l to access the state from layer l+2 at the same token position also naturally introduces a cycle. ARIA solves this by using detach() to access the previously computed state from the previous token; however, this also introduces a lag that may negate the usefulness of bridging in sequences with fast change.

Finally, the reduction in the number of hyperparameters to four also requires the identification of formal relationships between design choices and computational properties. For certain hyperparameters without formal relationships (e.g., θ, the EVM repair threshold), calibration on held-out data was also required, which introduces an empirical relationship that cannot be eliminated in the analysis.

Remedies:

Each of the challenges identified has a mitigation path through specific design choices that have been integrated into the ARIA architecture. For the cascade failure mode interactions, the mitigation path is the audit methodology. By cataloging all the failure modes prior to the initiation of the design process, the relationship between failure modes can be determined, and the order of activation can be planned for. This is achieved by the Class A/Class B partition, where structural failure modes are addressed prior to the optimization inefficiency upon which they depend.

The loss of global context, a problem of information degradation, occurs through the bottom-up compression problem. This problem is mitigated by the complementary nature of the three-stream approach, where Stream B addresses the problem of pyramid compression with causal constraints, and Stream C maintains global entity graphs through LET-guided attention, providing global information access through a different mechanism. PAT bypass ensures that precision-critical tokens such as numbers, dates, and identifiers are not lost through the averaging process, mitigating the most critical information degradation problem.

The LSM lag issue is addressed through the use of the per-layer learned scalar β_l, which is initialized to 0.01 and then annealed upwards throughout the training of Stage 3. This allows the model to adaptively calibrate the amount of cross-layer influence, minimizing dependence on lagged states when they are uninformative and maximizing dependence when the cross-layer context is beneficial.

The detach() function is used to break cycles, and the annealing schedule ensures that the model learns to effectively use the cross-layer information that is available to it.

The empirical dependency of θ is addressed through the formal characterization of the boundary conditions, which sets θ = 0.3 to trigger when the effective attention width drops below

approximately three keys (a near-degenerate distribution). This range is theoretically justified to be between 0.2 and 0.5. The formal derivation of k, W, and m from the complexity requirements provides three of the four hyperparameters without the need for empirical calibration, leaving only a single value to be determined.

Conclusion:

This paper proposes a novel, theoretically grounded multi-stream transformer architecture named ARIA, which solves all 18 previously identified failure modes of the prior sigmoid form of the mechanism. The six synergistic components of the ARIA architecture—namely, CR-RoPE, NSIG, CPC with CST and PAT, LET, EVM-PR,

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

and LSM—collectively convert each and every one of the previously identified instability modes into a theoretically grounded mode of stability. The proof of output boundedness for the NSIG component guarantees the numerical stability of the sought verification of CIO by ensuring ‖output_i‖_2 = √d for all sequence lengths N. This eliminates the 9,000-fold amplification present in the standard form of the sigmoid attention mechanism for sequence length N = 10,000 tokens. The guarantee of the CST component on the causal validity of the output also rules out the possibility of the hierarchical compression of the CIO output affecting the autoregressive generation. The guarantee of the LSM component on the cycle-free gradients also rules out the possibility of the support for multi-hop reasoning affecting the computation graph. EVM-based entropy monitoring is also capable of proactively detecting failures in the reasoning process that cannot be traced.

ARIA sets a new standard for the architectural soundness of long-context models, offering the first theoretical grounding for the Common Intent Orchestration (CIO) framework. By replacing heuristics with hard guarantees such as $O(N^{1.2})$ time complexity and a lower 1.3x KV cache overhead, ARIA ensures the reliability of multi-agent deployments on commodity hardware through a predictable four-stage training protocol that greatly simplifies the optimization process by reducing the hyperparameters of the model to a simple four-variable set. The main thesis of the research claims that the reliability of agentic governance hinges upon the reliability of the underlying model, and the authors' success in satisfying all ten requirements for a production-grade model at last closes the theoretical-practical divide between subquadratic efficiency and actual reliability. Though the current paper offers the mathematical roadmap, future work will focus on empirical verification of the latency of Causal Pyramid Compression and the interaction between the training schedule and the data curriculum, effectively turning the theoretical guarantee of ARIA into a global standard.

References:

1. *Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.*
2. *Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory- efficient exact attention with IO-awareness. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 35, 16344–16359.*
3. *Krishnan, A., & Mehta, D. (2025). Common intent orchestration: Dynamic semantic validation for secure multi-agent language model systems. In Proceedings of the IEEE/ACM International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS).*
4. *Ramapuram, J., et al. (2024). Theory, analysis, and best practices for sigmoid self-attention. arXiv preprint arXiv:2409.04431.*
5. *Rae, J., et al. (2020). Compressive transformers for long-range sequence modelling. In Proceedings of the International Conference on Learning Representations (ICLR).*
6. *Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2024). RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing, 568, 127063.*
7. *Sun, Y., Dong, L., Patra, B., Ma, S., Huang, S., Benhaim, A., Chaudhary, V., Song, X., & Wei, F. (2023). A*

ISSN: P-2455-0515
E- 2394-8450

Educreator Research Journal

OPEN ACCESS

Volume–XIII, Special Issues–I(b) Jan – Feb 2026

Original Research Article

length-extrapolatable transformer. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 14321–14337.

8. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.*
9. *Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 35.*
10. *Zaheer, M., et al. (2020). Big Bird: Transformers for longer sequences. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 33, 17283–17297.*

Cite This Article:

Singh L., Vishwakarma A., Yadav A. & Naikwadi A. (2026). *ARIA: Adaptive Rectified Integrated Attention — A Theoretically Grounded Multi-Stream Transformer Architecture Resolving Systemic Instabilities for Secure Common Intent Orchestration in Scalable Language Modeling.* **In Educreator Research Journal: Vol. XIII (Issue I),** pp. 29–36. **Doi:** https://doi.org/10.5281/zenodo.19915867