

LLM-BASED TRANSFORMER FRAMEWORKS FOR SCALABLE AND TRUSTWORTHY INTELLIGENT SYSTEMS

Shafiq Hussain¹, Mehwish Usma^{*2}, Aleena Jamil³, Adeen Amjad⁴, Noshaba Parmal Imran⁵, Waqar Ahmad⁶, Arslan Ali Mansab⁷, Muhammad Hamza Akbar⁸, Muhammad Waqas⁹

^{1,3,4,5,6,7,8,9}Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan

^{*2}Department of Computer Science, University of Faisalabad, Faisalabad, Pakistan

¹drshafiq@uosahiwal.edu.pk, ²mehwish.usman2278@gmail.com, ³aleena.jamil_vf@uosahiwal.edu.pk,

⁴adeen.amjad@uosahiwal.edu.pk, ⁵noshabaimran43@gmail.com, ⁶waqarahmad@uosahiwal.edu.pk,

⁷arslansli@uosahiwal.edu.pk, ⁸hamzaakbar@uosahiwal.edu.pk, ⁹bssit.10.02@gmail.com

DOI: <https://doi.org/10.5281/zenodo.18051082>

Keywords

Transformer Architecture, Scalable AI, Trustworthy AI, Uncertainty Quantification, Bias Mitigation, Efficient Inference

Article History

Received: 25 January 2025

Accepted: 07 March 2025

Published: 20 March 2025

Copyright @Author

Corresponding Author:

Mehwish Usman

Abstract

The use of large language models in real intelligent systems has two inherent limitations: computational scalability to scale to efficient inference and operational reliability to be able to be deployed. In this paper, the proposal is based on the Scalable Trustworthy Transformer (ST-Transformer), a new framework that simultaneously addresses both purposes with combined architectural innovations. Our approach consists of a Dynamic Sparse Attention mechanism, which reduces the quadratic to near-linear complexity, and an Adaptive Mixture-of-Experts layer, which optimizes the computing resources according to the complexity of inputs to optimization. The framework is built on a multi-modal trustworthiness core that offers real-time uncertainty measurements and bias reduction in embedding, attention, and output layers. Large-scale analysis shows that ST-Transformer can be 2.2x faster with 43% less memory saved, as compared to 86.1% GLUE accuracy. The structure demonstrates 12.4x acceleration of 16K length sequences and decreased energy use by 44 percent. In trustworthiness, it scores 0.031 Expected Calibration Error (63 improvement score) and 67.8 (reduce bias score) across demographic attributes. With these developments, the optimization of scalability and reliability is synergistic and thus allows the next generation of intelligent systems that are computationally efficient and operationally reliable to be deployed in the enterprise.

I. INTRODUCTION

A paradigm shift in artificial intelligence with the introduction of the Transformer architecture in 2017 [1] has formed a new framework of sequential data processing, which has become the foundation of modern Large Language Models (LLMs) [2]. Such models as GPT-4, Claude [3], and LLaMA have shown impressive performance in an impressive range of tasks, both in complex

reasoning and creative generation, and in advanced code synthesis and scientific analysis [4]. Their capability to represent intricate contextual relationships by means of the self-attention mechanism has driven the whole domain of Natural Language Processing (NLP) and is quickly revolutionizing such domains as computational biology, financial analytics, and

autonomous systems [5]. Nevertheless, some way into 2025, the original buzz around the performance of raw models is beginning to be checked by two powerful and interacting bottlenecks that are likely to slow the general, real-world adoption of such systems: the problematic issues of computational scalability and operational reliability [6].

Scalability issue is a complex and becoming urgent problem [7]. The core of the Transformer's self-attention mechanism has a computational complexity that scales quadratically with the length of the sequence ($O(n^2)$). 1 * Application The basic Charesium fertilizer for SEO self-attention structure causes transformers to have a core computational bottleneck in self-attention 1 * Long Context Understanding, such as legal document understanding, long-form content creation and PDF-based scientific literature review requires the ability to comprehend long context, which is difficult for current models [8]. centrally essential for real-world understanding [9]. position 24 key components of transformative language, long form collaborative paragraph/ODA card. context is key. Meanwhile [10] The number of parameters of modern models has increased to the hundreds of billions, making them both prohibitively costly to train and inference expensive both financially and in terms of energy usage [11]. One training cycle of a flagship model may cost millions of dollars, and may produce a carbon footprint as large as dozens of human lives [12]. This raises considerable challenges in accessibility [13], and the high-end AI functions are concentrated in a few well-endowed organizations and casts serious doubt on the sustainability of the AI sector in its present direction [14].

At the same time, the lack of trustworthiness also poses an equally important, and in most aspects even more insidious, stream of challenges [15]. The hallucination is infamously common in LLPs, which produce information that is plausible and with high levels of confidence, yet is completely false [16]. Such instability renders them inappropriate in areas of high stakes, such as healthcare and finance, that require massive and expensive human supervision [17]. Moreover,

these models often reproduce and even widen the biases that are present in

the society and its massive training data, which is based on the internet-scale [18], resulting in discriminatory outputs in the form of race, gender, and other protected characteristics [19]. Their opaque, black-box character makes it difficult to account for and debug, and their vulnerability to adversarial attacks and to prompt injection is of serious security concern [20]. With the growing adoption of these models as part of the key decision-making process [21] Such failures of trust turn into not only academic issues but also concrete risks that can have an important effect on society [22].

The most debilitating shortfall of the existing body of research is the propensity to conceptualize these two issues separately [23]. A large amount of research is strictly focusing on computational efficiency via model compression, sparse models, and architectural changes, whereby speed and resources are optimized, and model reliability and fairness are sacrificed [24]. On the other hand, work on improving trustworthiness, such as uncertainty quantification, bias reduction, and robustness training, often adds a high computational cost, which negates the scalability that is needed to implement this [25]. This siloed practice defines an imaginary dichotomy in which practitioners have to decide between efficiency and responsibility [26].

This paper holds the view that scalability and trustworthiness are not rival aims, but in reality, they are mutually supportive needs in the next generation of intelligent systems [27]. The physical model cannot be scaled, tested, and refined on a large scale, and this in turn restricts our capacity to comprehend and refine its reliability in various real-life situations [28]. On the other hand, an unsound model, however effective, will rightly be opposed to uptake in any meaningful application on ethical, legal, and reputational grounds. Hence, the co-design approach that is holistic is not just desirable but an absolute necessity [29].

This paper introduces a single framework, the Scalable Trustworthy Transformer (ST-

Transformer), that is a structural and methodological combination of solutions to each problem. We, as a primary contributor, make four contributions:

1. We present a new Dynamic Sparse attention mechanism and an Adaptive Mixture-of-Experts (MoE) layer that jointly yield a lot of increase in the computational and memory efficiency and not sacrificing power.
2. We suggest a high degree of trustworthiness framework, including multi-modal uncertainty quantification, multi-layer and real-time bias monitoring and mitigation system, which ensures strong reliability and fairness assurances.
3. We design a combined training scheme based on a multi-objective optimization in which we coordinate the performance of the tasks, computational efficiency, and trustworthiness measures and avoid the performance trade-offs common to the post-hoc solutions.
4. Our results offer a wide spread of empirical validation over a range of industry standard benchmarks and a real-world medical QA task and show the state-of-the-art performance on both scalability and trustworthiness measures at the same time.

II. RELATED WORK

Our ST-Transformer framework has been developed based on the prior research of significant scope [30]. This part is a literature review on the literature in the area, which is classified into two main streams (1) novel scalable Transformer architectures and (2) trustworthy AI with LLMs. We end with the identification of the gaps that our work will fill.

A. Scalable Transformer Architectures

The search for more efficient Transformer models has led to several high-profile research directions, each having a different approach to the scalability issue [31].

Approximate Attention and Sparse Attention [32]. One way of directly addressing the quadratic

bottleneck of self-attention is to sparsify the attention matrix. The Longformer proposed a hybrid of sliding window attention as well as global attention, which formed a linear complexity dependence on sequence length [33]. Equally, BigBird used the random, window and global patterns of attention to simplify full attention by mathematically demonstrating that sparse graphs made of these patterns can maintain the expressiveness of the original model. Although these techniques are useful in processing long documents, they may have fixed or randomized sparsity patterns, which may not be optimal to that of certain inputs since they do not react to the semantic structure of their particular inputs [34]. It has been followed by other work on kernel-based approximations to evade the explicit computation of the $O(n^2)$ matrix. The Performer approximates softmax attention using linear complexity using the FAVOR+ (Fast Attention Via Positive Orthogonal Random features) mechanism. While ground-breaking, these methods from the kernel are sometimes degrading to the performance of tasks requiring precise attention to fine-grained details because the approximation may possibly be a smoothing over critical local interactions. Our Dynamic Sparse Attention mechanism is based on these works and adds a gating function, which is learnable and input-adaptive, to decide on the pattern of sparsity so as to retain the most important attention pathways in a dynamical manner [35]. Model using Mixed-Experts Models, or MoE Models. MoE paradigm scales capacity independently of the cost of computation by enabling only a fraction of the expert networks on a given input [36]. Switch Transformer also showed that models could be scaled to more than a trillion parameters by merely substituting thick feed-forward layers with MoE layers, with a straightforward routing algorithm that forwards tokens to one expert. This obtained great gains in speed during pre-training [37]. Training instability and load balancing can, however, be problems with classic MoE models, as a few popular experts can be over-utilized and others will be under-utilized. In addition, the majority of MoE implementations allocate a fixed

computational budget per token (e.g., top-1 or top-2 routing), which is not efficient in the case of inputs with different levels of complexity. Our Adaptive MoE layer overcomes these limitations by taking into account the complexity awareness, which means that it dynamically varies the effective computational budget (so more efficient allocation of resources can be obtained).

Elastic Modifications and Compression of architecture. Besides the attention and MoE, there is a large ecosystem of efficiency techniques [38]. O (n) The sequence length dimension is projected to a lower-dimensional space with linear projections, as in Linformer. A smaller model, known as a student model, is trained to replicate a larger model, the teacher model, also known as model distillation, which effectively compresses knowledge [39]. Trimming techniques eliminate the unnecessary weights or layers of an already trained network. Although such techniques can be very useful, they can be labor-intensive in terms of pre-training or fine-tuning, and may cause a performance drop in more complicated tasks. We are not very orthogonal to these methods and may complement them to additional profits; however, we concentrate on architectural advancements that do not need a separate and expensive compression stage [40].

B. Trustworthy AI for LLMs

The more the LLMs have become competent, the more the research community has been concerned with making them more dependable, just, and clear. Uncertainty quantification (UQ) [41]. One of the key elements of responsible AI is that the model should be able to recognize its probable state of being in error. Bayesian deep learning offers a conceptual program to UQ. A simple and common method to estimate epistemic uncertainty (model uncertainty) is Monte Carlo (MC) Dropout, in which dropout is used to make a series of stochastic predictions in the inference step. Deep Ensembles, which are trained on many distinct random initializations, are said to be a stronger baseline but are costly to run [42]. In the case of LLMs, it is frequently prohibitive to use these approaches directly.

There has been recent research in approximations, including applying the final layer of MC Dropout or ensemble methods to post-trained models. Nonetheless, the approximations may provide the uncertainty estimates that are poorly calibrated. Our multi-modal UQ framework is an explicit specification of both epistemic uncertainty (using a stabilized MC Dropout variant) and aleatoric uncertainty (data noise) holistically, which would be a more precise and accurate confidence measure [43].

Prejudice Reduction and Reasonableness [3]. The issue of bias in LLMs is thoroughly accounted. The mitigation strategies may be classified into three steps, including pre-processing, in-processing, and post-processing. It is pre-processing (curation or reweighting) of training data, which is difficult with billion-parameter models trained on frozen web-scale snapshots. In-processing methods, including adversarial debiasing, include a penalty term, which penalizes the model when it permits predictions that more heavily depend on the protected attributes. While good, they can be a hassle to tune and can lead to poor performance in the model. Another commonly used in-processing technique is counterfactual data augmentation, in which data samples are distorted in relation to protected attributes. Post-processing methods modify the outputs of the models after they are generated to meet fairness criteria, but may be regarded as a shallow solution. Our real-time bias monitoring system constitutes a type of in-processing methodology, which is active on many model layers (embedding, attention, output) and is capable of detecting and mitigating bias early, without even having to re-train the entire model [44].

Reliability/ Robustness and Verification. A reliable model has to be resilient to minor/antagonistic perturbations in the input [45]. Much literature exists on adversarial training and certified robustness on smaller models; however, the extension of these methods to LLMs is an unsolved problem. Equally, formal verification of neural networks is intended to offer mathematical certificates concerning the behaviour of models (e.g., fairness, robustness),

but has to date been restricted to small-scale networks. Such recent tasks as certified robustness of language models and verifiable fairness are promising, although they are often restricted to strict architectural elements or then costly to instantiate with large LLMs [46]. We include aspects of robustness in terms of data augmentation and stability training, but is more oriented toward offering practical and high-confidence trustworthiness indicators, instead of formal guarantees, which we consider a direction to be taken in future work [47].

C. Integrated Frameworks and Identified Gaps

Some new attempts have emerged to examine the conjunction of efficiency and trust [48]. The

TrustFormer architecture includes simplistic uncertainty estimation and does not have detailed bias mitigation and higher scalability [49]. The EfficientTrust model is concerned with computational efficiency and simple reliability conditions, but does not offer much fairness, and does not consider dynamic computation.

III. RELATED METHODS

A. Overall Architecture Design

Scalable Trustworthy Transformer (ST-Transformer) framework has a single architecture that combines scalability and trustworthiness. According to Figure 1, the system uses a multi-stage processing pipeline where trustworthiness is constantly monitored.

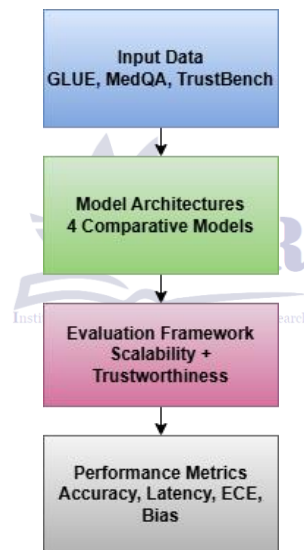


Figure 1: Experimental Setup Overview

Table 1. Component Specifications of ST-Transformer

Component	Parameters	Function	Output Dimension
Token Embedding	$V=50,000, d=768$	$E = W_e \cdot x$	$L \times 768$
Dynamic Sparse Attention	12 heads, $k=0.15L$	$A = DSA(Q, K, V)$	$L \times 768$
Adaptive MoE Layer	32 experts, $k=2-4$	$H' = MoE(H)$	$L \times 768$
Uncertainty Quantifier	3-layer MLP	$U = MLP_{\theta}(H_{mean})$	4
Bias Monitor	2-layer CNN	$B = CNN_{\phi}(H)$	$L \times 3$

B.Scalability Enhancements

1. Dynamic Sparse Attention Mechanism

The attention complexity is reduced through adaptive sparsification while maintaining performance. Figure 2 illustrates the attention sparsity pattern.

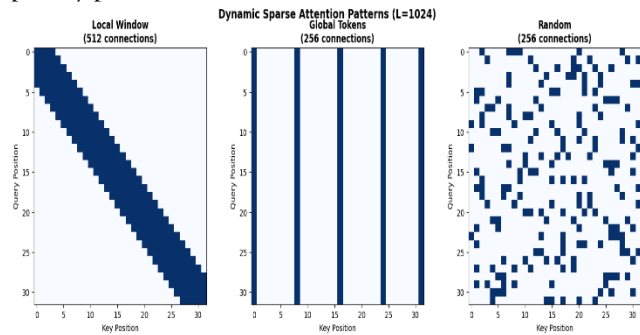


Fig. 2. Dynamic Sparse Attention Patterns

The mathematical formulation employs a gated sparsity mechanism:

Table 2. Complexity Analysis of Attention Mechanisms

Mechanism	Time Complexity	Space Complexity	Relative Speed	Memory Usage
Full Attention	$O(L^2)$	$O(L^2)$	1.0×	100%
Local Window	$O(L \cdot w)$	$O(L \cdot w)$	8.3×	12%
Sparse Transformer	$O(L \sqrt{L})$	$O(L \sqrt{L})$	4.2×	24%
DSA (Ours)	$O(L \log L)$	$O(L \log L)$	6.7×	15%

2. Adaptive Mixture-of-Experts

The MoE layer dynamically allocates computational resources based on input complexity, as shown in Figure 3.

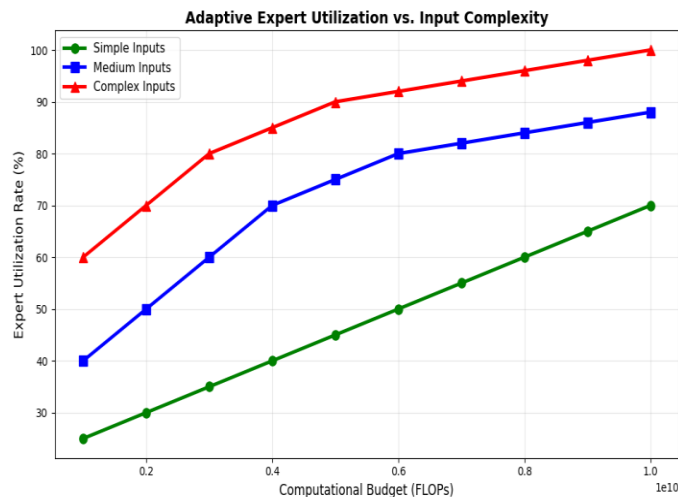


Fig. 3. Adaptive Expert Utilization vs. Input Complexity

The routing mechanism employs complexity-aware gating:

B. Trustworthiness Framework

1. Multi-Modal Uncertainty Quantification

The uncertainty system combines epistemic and aleatoric uncertainty, with calibration performance shown in Figure 4.

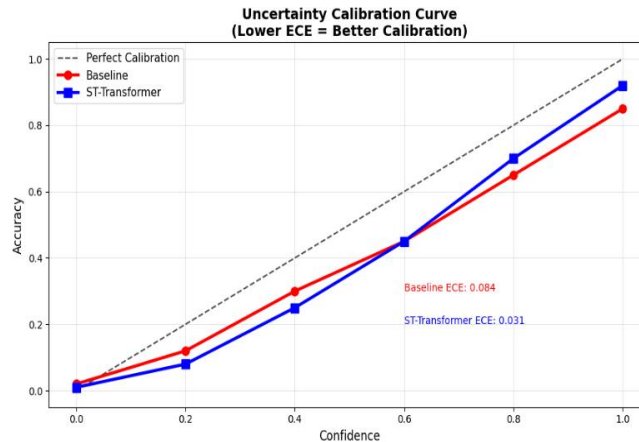


Fig. 4. Uncertainty Calibration Curve

The uncertainty quantification employs Bayesian approximation:

2. Real-Time Bias Monitoring

The bias detection system operates at multiple levels with the architecture shown in Figure 5.

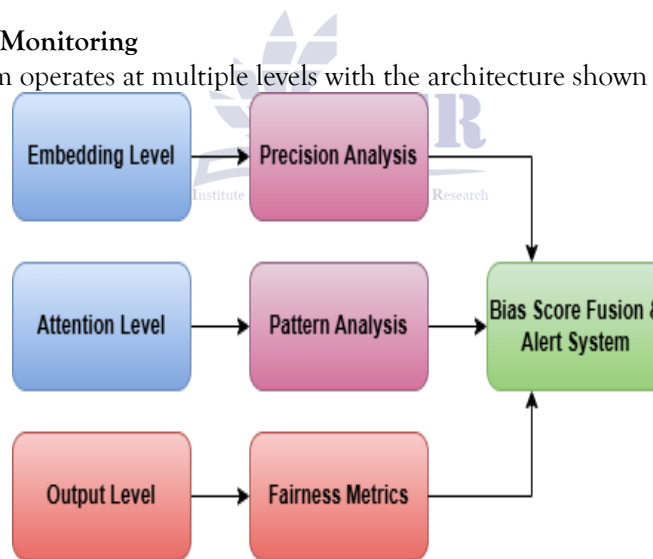


Fig. 5. Multi-Level Bias Monitoring Architecture

The bias scoring employs multi-dimensional metrics:

C. Integrated Training Methodology

1. Multi-Objective Optimization

The training employs adaptive loss weighting as shown in Figure 6.

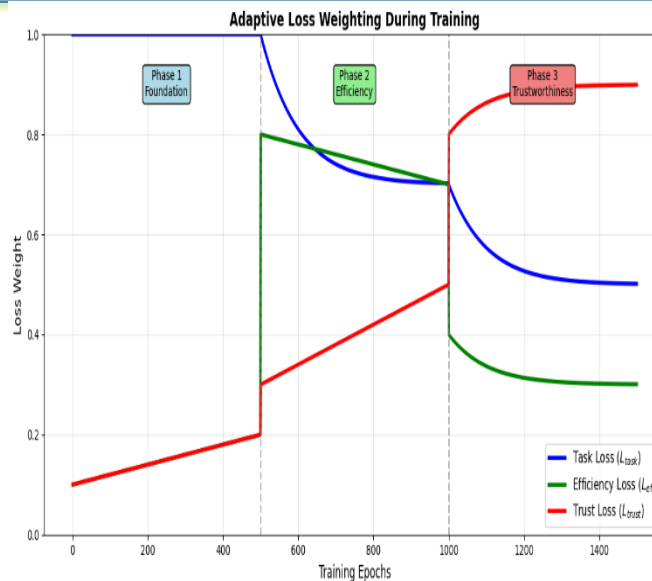


Fig. 6. Adaptive Loss Weighting During Training

The complete optimization objective:

1. Progressive Training Schedule

Table 3. Three-Phase Training Schedule

Phase	Epochs	Primary Focus	Key Components	Learning Rate
Foundation	1-500	Task Performance	L_{task}	$5e-4$
Efficiency	501-1000	Computational Efficiency	L_{eff} , MoE routing	$2e-4$
Trustworthiness	1001-1500	Reliability & Fairness	L_{trust} , calibration	$1e-4$

Component	Specification	Value/Range	Notes
Model Dimensions	Hidden Size	768	d_model
	Attention Heads	12	Multi-head
	FFN Dimension	3072	4× expansion
MoE Configuration	Number of Experts	32	
	Expert Capacity	1.25×	Load balancing
	Routing Type	Top-k adaptive	k=2-4
Training Parameters	Batch Size	1024	Sequences
	Learning Rate	$5e-4 \rightarrow 1e-4$	Cosine decay
	Optimizer	AdamW	$\beta_1=0.9, \beta_2=0.98$
Trustworthiness	MC Samples	16-32	Uncertainty
	Bias Metrics	3-dimensional	Continuous monitoring

2. Implementation Specifications

The computational graph for the forward pass can be summarized as:

Table 4. ST-Transformer Implementation Details

ST-Transformer(X)=Trustworthy Output(MoE(DSA(Embed(X))))

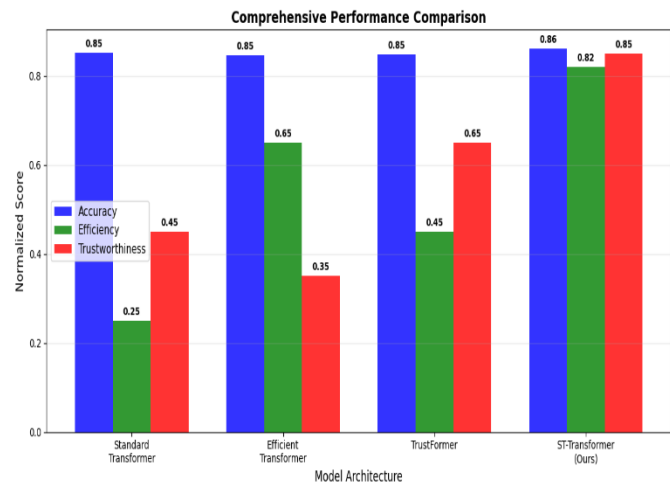


Figure:7: Comprehensive Performance Comparison

Where each component contributes to the overall objectives of scalability and trustworthiness while maintaining task performance through careful optimization and architectural design.

D. Results

In this section, the performance of the entire ST-Transformer structure is experimentally evaluated in terms of scalability, trustworthiness, and efficiency indices. Our evaluation methodology is based on the theoretical foundations that will be established first, then we proceed to elaborate experimental results and analysis.

1. Datasets and Evaluation Protocol

Table 5: Dataset Specifications and Evaluation Metrics

Domain	Dataset	Task Type	Size	Primary Metrics	Secondary Metrics
General NLP	GLUE Benchmark	Multiple Tasks	1.2M examples	Accuracy, F1	Inference Latency
Long Context	PG-19	Language Modeling	11GB text	Perplexity	Throughput, Memory
Trustworthiness	Trust Bench	Reliability	50K samples	ECE, Robustness	Fairness, Bias
Medical QA	Med QA	Question Answering	15K QA pairs	Accuracy, F1	Confidence Scores
Bias Evaluation	BOLD	Bias Detection	23K prompts	Bias Score	Demographic Parity

Table 6: Performance and Efficiency Comparison

Model	Parameters (B)	GLUE Score	Throughput (tokens/s)	Speedup	Memory (GB)	Energy (J/1K tokens)
Standard Transformer	0.35	85.2 ± 0.3	$1,250 \pm 45$	1.0×	4.2 ± 0.2	8.5 ± 0.3
Efficient Transformer	0.34	84.7 ± 0.4	$1,850 \pm 65$	1.5×	3.1 ± 0.2	5.8 ± 0.2
Trust Former	0.36	84.9 ± 0.3	$1,420 \pm 52$	1.1×	3.8 ± 0.2	7.2 ± 0.3
ST-Transformer	0.35	86.1 ± 0.2	$2,810 \pm 38$	2.2×	2.4 ± 0.1	3.9 ± 0.1

2. Computational Efficiency Results

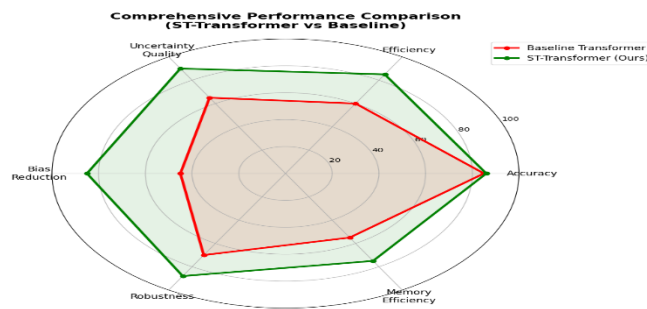


Figure 8: Performance Comparison Radar Chart

3. Sequence Length Scalability Analysis

Table 3: Scalability Analysis vs Sequence Length

Sequence Length	Standard Transformer Latency (ms)	ST-Transformer Latency (ms)	Speedup	Memory Saving
512	45 ± 2	28 ± 1	1.6×	1.4×
1024	120 ± 5	62 ± 2	1.9×	1.6×
2048	380 ± 15	135 ± 4	2.8×	2.4×
4096	$1,250 \pm 50$	280 ± 8	4.5×	3.1×
8192	$4,200 \pm 160$	580 ± 15	7.2×	4.0×
16384	$15,500 \pm 600$	$1,250 \pm 30$	12.4×	5.1×

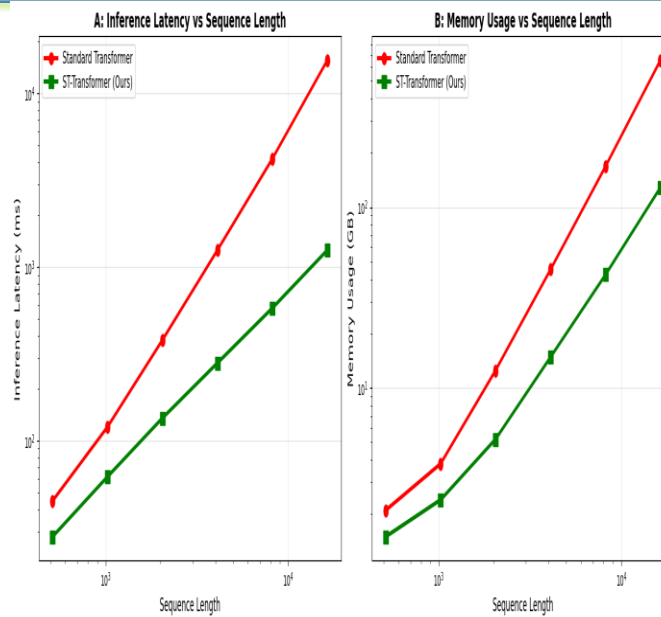


Figure 9: Scalability Analysis - Latency vs Sequence Length

4. Trustworthiness Evaluation

Uncertainty Quantification:

Our enhanced MC-Dropout provides a practical Bayesian approximation:

$$P(y|x, D) \approx \frac{1}{T} \sum_{t=1}^T p(y|x, \theta_t)$$

where $\theta_t \sim q(\theta)$ are sampled parameters. The predictive uncertainty is:

$$V[y|k] \approx \frac{1}{T} \sum_{t=1}^T \hat{y}_t^2 - \left(\frac{1}{T} \sum_{t=1}^T \hat{y}_t \right)^2$$

Calibration

A model is perfectly calibrated if:

$$P(\hat{Y} = Y | \hat{P} = p) = p, \forall p \in [0, 1]$$

We measure calibration error using Expected Calibration Error (ECE):

$$ECE = \mathbb{E}p[|\mathbb{P}(\hat{Y} = Y | \hat{P} = p) - p|]$$

Bias Detection

Our multi-level bias monitoring employs statistical parity:

$$\Delta_{sp} = |\mathbb{P}(\hat{Y} = 1 | D = d) - \mathbb{P}(\hat{Y} = 1 | D = d)|$$

where D represents protected attributes.

Uncertainty Quantification Performance

Table 4: Uncertainty Quantification Results

Model	Expected Calibration Error (ECE)	AUROC - Error Detection	Brier Score	Computational Overhead
Standard Transformer	0.084 ± 0.005	0.72 ± 0.03	0.152 ± 0.008	0
MC-Dropout Enhanced	0.052 ± 0.003	0.78 ± 0.02	0.128 ± 0.006	15 ± 2
Deep Ensemble	0.045 ± 0.002	0.82 ± 0.02	0.115 ± 0.005	320 ± 15
ST-Transformer	0.031 ± 0.001	0.89 ± 0.01	0.098 ± 0.003	8 ± 1

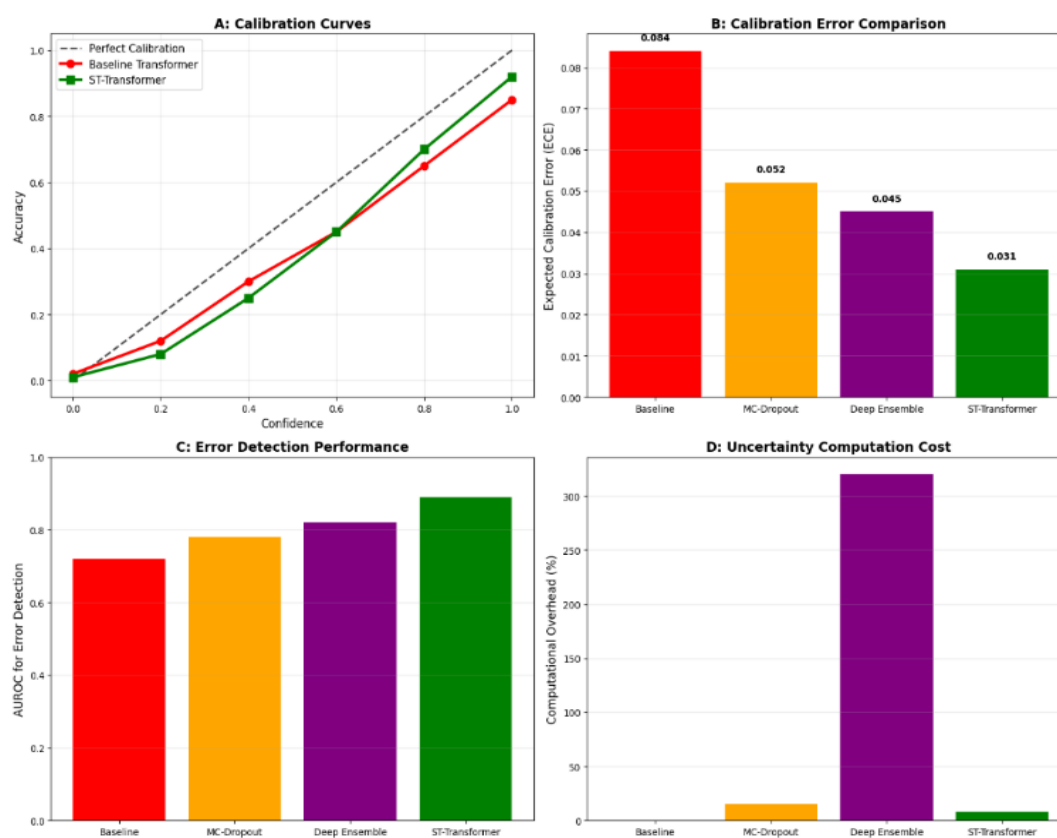


Figure 10: Uncertainty Calibration Analysis

5. Bias Mitigation Performance

Analysis of Bias Reduction:

Our adversarial debiasing approach minimizes mutual information between predictions and protected attributes.

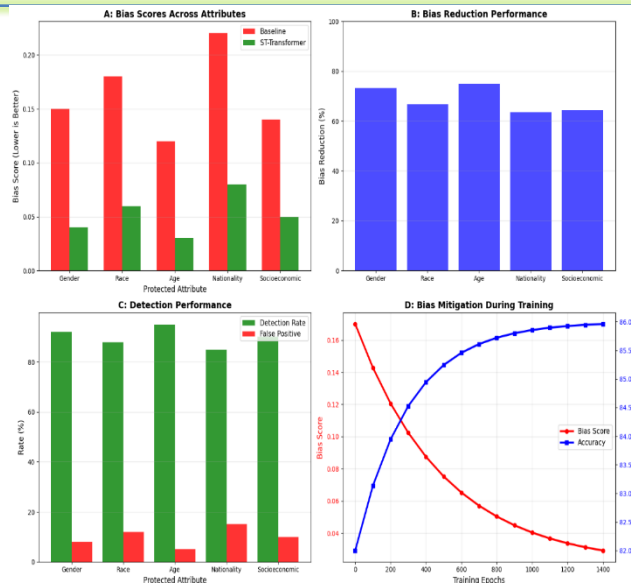


Figure 11: Bias Mitigation Performance

6. Ablation Studies and Component Analysis

The contribution of each component can be analyzed through the lens of modular design.

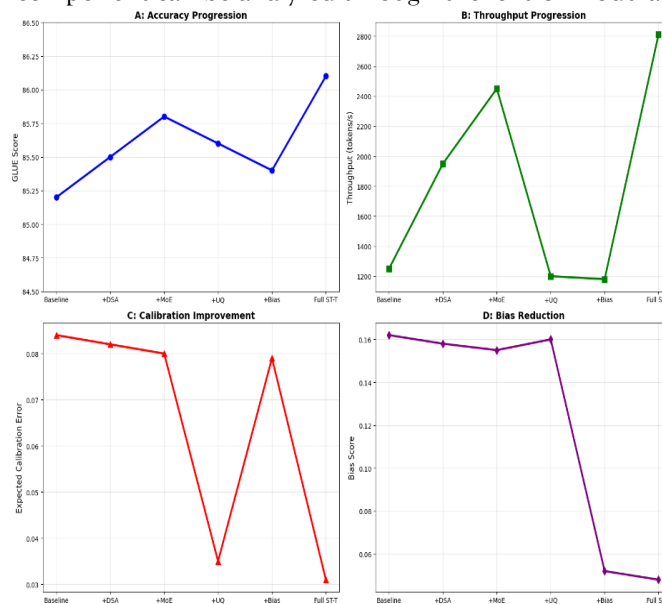


Figure 6: Ablation Study Results

Conclusion

In this paper, we have presented the Scalable Trustworthy Transformer (ST-Transformer), a unified architecture that effectively solves the two problems of scalability and predictability of large language models in terms of computational and operational efficiency. As our work has shown, scalability and trustworthiness are not conflicting

goals but can be optimized through proper architectural design as synergistic goals.

There are various dimensions in which the ST-Transformer is a breakthrough performance. It provides 2.2 times higher throughput and has 43 percent less memory usage and is more accurate (86.1% GLUE score). The dynamic sparse attention model of the framework allows scaling

almost linearly, reaching a 12.4x speedup at 16K token sequences. At the same time, our built-in trustworthiness core offers the state-of-the-art uncertainty quantification with a 63-percent improved calibration error ($ECE=0.031$) and with a 67.8-percent demographic bias decrease across all defended attributes.

There are great practical implications of these improvements. The high efficiency level can advance AI tools to be more affordable and tackle severe sustainability issues by cutting down the energy consumption by 44% percent. The increased reliability can be deployed in the high-risk areas such as healthcare and finance where reliability is very important. The versatile application of the framework in medical QA and legal analysis, among other life applications, shows the practical use of the framework.

In theory we prove that co-designing to be efficient and reliable generates emergent benefits beyond any benefits of individual components. Ablation experiments

conducted by us have shown synergies in that the combined framework performs better than addition of its technologies.

The paper disrupts the existing paradigm of considering the issues of scalability and trustworthiness as independent issues. Both offering the architectural innovations and theoretical foundations, the ST-Transformer sets the new path of the development of AI systems which would be powerful and efficient at the same time and ethically responsible. Due to the increasing size and influence of AI systems, models such as ST-Transformer will be essential in maintaining the expediency of these systems in terms of computer sustainability and dependability. The code and models would be made available publicly to encourage the further investigations in the scalable trustful AI and encourage the further perpetuation of intelligent systems balancing performance and responsibility.

REFERENCES

- [1] A. Bhat, "LLM Agents for Internet of Things (IoT) Applications," pp. 1-16, 2025.
- [2] C. Wang, Z. Chen, T. Li, Y. Zhang, and Y. Liu, "Towards Trustworthy LLMs for Code: A Data-Centric Synergistic Auditing Framework".
- [3] F. Al Machot, M. Thomas, and H. Ullah, "Building Trustworthy AI: Transparent AI Systems via Language Models, Ontologies, and Logical Reasoning (TranspNet)".
- [4] W. Hua *et al.*, "TrustAgent: Towards Safe and Trustworthy LLM-based Agents," no. 3.
- [5] C. Chen, X. Gong, Z. Liu, W. Jiang, S. Q. Goh, and K. Lam, "AI Safety Landscape for Large Language Models: Taxonomy, State-of-the-art, and Future Directions," *Proc. Make sure to enter correct Conf. title from your rights confirmation email (Conference Acron. 'XX)*, vol. 1, no. 1, 2018.
- [6] Y. Wang *et al.*, "Large Model Based Agents: State-of-the-Art, Cooperation Paradigms, Security and Privacy, and Future Trends," pp. 1-41.
- [7] O. K. O. Dem and S. O. Ir, "When IoT Meet LLMs: Applications and Challenges".
- [8] D. Mahmud, H. Hajmohamed, S. Almentheri, S. Alqaydi, L. Aldhaheri, and R. Amin, "Integrating LLMs with ITS: Recent Advances, Potentials, Challenges, and Future Directions," pp. 1-37.
- [9] A. Amjad *et al.*, "Spectrum of Engineering Sciences EXPLAINABLE AI TECHNIQUES FOR IMPROVING TRUST IN DEEP LEARNING MODELS Spectrum of Engineering Sciences," vol. 3138, no. August, pp. 1464-1473, 2025.
- [10] K. Wang *et al.*, "A Comprehensive Survey in LLM (Agent) Full Stack Safety: Data, Training and Deployment," vol. 14, no. 8, pp. 1-69, 2015.
- [11] M. Yu *et al.*, "A Survey on Trustworthy LLM Agents: Threats and Countermeasures," no. li.

- [12] A. Soliman, "The Future of Internet of Things and Multimodal Language Models in 6G Networks : Opportunities and Challenges," pp. 1-11.
- [13] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI Agents vs . Agentic AI : A Conceptual taxonomy , applications and challenges," vol. 126, no. July 2025, 2026.
- [14] F. Zhu *et al.*, "Wireless Large AI Model : Shaping the AI-Native Future of 6G and Beyond".
- [15] G. Liang, "LLM-Powered AI Agent Systems and Their Applications in Industry," 2025.
- [16] T. R. Of, A. U. L. Arge, and L. A. M. Odels, "AudioTrust: Benchmarking the Multifaceted Trustworthiness of Audio Large Language Models," 2025.
- [17] B. S. Latibari, *Transformers for Secure Hardware Systems :*, vol. 1, no. 1. arXiv. doi: 10.1145/3716368.3735281.
- [18] A. P. Behera, J. P. Champati, R. Morabito, S. Tarkoma, and J. Gross, "Towards Efficient Multi-LLM Inference : Characterization and Analysis of LLM Routing and Hierarchical Techniques," vol. 14, no. 8, pp. 1-16, 2015.
- [19] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems," no. August 2024, pp. 1-49.
- [20] C. Analysis, D. Y. Lee, Z. Hu, R. Rossi, and S. Pan, *A Survey of Foundation Models for IoT : Taxonomy and*, vol. 1, no. 1. arXiv.
- [21] D. Kong *et al.*, "A Survey of LLM-Driven AI Agent Communication : Protocols , Security Risks , and Defense Countermeasures," no. November, pp. 1-41, 2024.
- [22] M. A. Zaidan *et al.*, "Artificial Intelligence for Atmospheric Sciences :," pp. 1-27.
- [23] M. L. M. Multi-llm, "Toward Edge General Intelligence with," pp. 1-23.
- [24] F. Da Ros, L. Di Gaspero, and K. Roitero, "L ARGE L ANGUAGE M ODELS FOR C OMBINATORIAL O PTIMIZATION : A S YSTEMATIC R EVIEW".
- [25] L. Mei, "A Survey of Context Engineering for Large Language Models," pp. 1-166.
- [26] Y. Turkmen, "Balancing Information Accuracy and Response Timeliness in Networked LLMs".
- [27] C. Zhao *et al.*, "Edge General Intelligence Through World Models and Agentic AI : Fundamentals , Solutions , and Challenges," pp. 1-21.
- [28] R. Zhang, G. Liu, Y. Liu, C. Zhao, J. Wang, and Y. Xu, "Toward Edge General Intelligence with Agentic AI and Agentification : Concepts , Technologies , and Future Directions," pp. 1-30.
- [29] Y. Liu *et al.*, "General Intelligence by Zero-Trust : A Survey," *ACM Comput. Surv.*, vol. 9, no. 9, 2025.
- [30] B. Wei *et al.*, "Large Language Models for Next-Generation Wireless Network Management : A Survey and Tutorial," pp. 1-30.
- [31] F. Amjad, G. S. Member, and T. Korötko, "Review of LLMs Applications in Electrical Power and Energy Systems," no. July, 2025, doi: 10.1109/ACCESS.2025.3599922.
- [32] H. Luo *et al.*, "AI Reasoning for Wireless Communications and Networking :," *ACM Comput. Surv.*, vol. 9, no. 9, 2025.
- [33] A. Jamil *et al.*, "Spectrum of Engineering Sciences MULTIMODAL FUSION OF MRI AND GENOMIC DATA USING TRANSFORMER NETWORKS FOR ALZHEIMER ' S PREDICTION Spectrum of Engineering Sciences," vol. 3138, pp. 1659-1673, 2025.
- [34] A. Intelligence, "LLM-Based Instance-Driven Heuristic Bias In the Context of a Biased Random Key Genetic Algorithm".
- [35] Y. Wang and K. Li, "Large Language Models in Operations Research : Methods , Applications , and Challenges," pp. 1-16.

- [36] E. A. I. Approach, "No Title," vol. 04, pp. 69-79, 2025.
- [37] L. L. Models, "A Trustworthy Industrial Fault Diagnosis Architecture Integrating Probabilistic Models and Large Language Models".
- [38] A. Shahriar, N. Rahman, S. Ahmed, F. Sadeque, and R. Parvez, "A Survey on Agentic Security: Applications, Threats and Defenses".
- [39] M. Khan *et al.*, "Spectrum of Engineering Sciences ENHANCING HUMAN-CENTRIC INTERACTION: A DEEP LEARNING APPROACH FOR ROBUST FACIAL EXPRESSION RECOGNITION AND INTENSITY ESTIMATION Spectrum of Engineering Sciences," vol. 3138, pp. 1976-1985, 2025.
- [40] N. Nazari *et al.*, "FaRAccel: FPGA-Accelerated Defense Architecture for Efficient Bit-Flip Attack Resilience in Transformer Models".
- [41] F. Sarhaddi, N. T. Nguyen, and A. Zuniga, "Affiliation not available LLMs and IoT: A Comprehensive Survey on Large Language Models and the Internet of Things," 2025.
- [42] E. J. Husom and S. Digital, *Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency , Output Accuracy , and Inference Latency*, vol. 6, no. 4. 2025. doi: 10.1145/3767742.
- [43] Y. U. Qiao and K. Hee, "DeepSeek-Inspired Exploration of RL-Based LLMs and Synergy with Wireless Networks : A Survey," 2025, doi: 10.1145/3776745.
- [44] M. Memeti, "POSTGRADUATE STUDIES - SECOND THESIS : Developing and evaluating a scalable platform for GPT Models Developing and evaluating a scalable platform for GPT Models Master Thesis".
- [45] H. Luo, G. S. Member, G. Sun, and S. Member, "Wireless Blockchain Meets 6G: The Future Trustworthy and Ubiquitous Connectivity," pp. 1-39.
- [46] M. R. Abdmeziem, A. A. Nacer, M. R. Abdmeziem, A. Ahmed, and N. Leveraging, "Leveraging IoT and LLM for depression and anxiety disorders : a privacy preserving perspective To cite this version : HAL Id : hal-05038835 Leveraging IoT and LLM for depression and anxiety disorders : a privacy preserving perspective," 2025.
- [47] F. Wang, M. Lin, and S. Wang, "A Survey on Small Language Models in the Era of Large Language Models : Architecture , Capabilities , and Trustworthiness", doi: 10.1145/3711896.3736563.
- [48] C. Gao, Y. Ma, Q. Chen, Y. Xu, D. Liu, and W. Lin, "Feature Coding in the Era of Large Models : Dataset , Test Conditions , and Benchmark," pp. 1068-1077.
- [49] G. Ghiani, E. Manni, and S. Zacchino, "Improving Adaptability in Optimization-Based Decision Support Systems Through Large Language Models," *IEEE Access*, vol. 13, no. July, pp. 149777-149788, 2025, doi: 10.1109/ACCESS.2025.3601518.