Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 4, 2025

A TRUST-AWARE TRANSFORMER ARCHITECTURE WITH UNCERTAINTY ESTIMATION FOR HIGH-STAKES LANGUAGE-BASED DECISION SYSTEMS

Shafiq Hussain[1], Muhammad Arman[*2], Adeen Amjad[3], Aleena Jamil[4], Waqar Ahmad[5], Arslan Ali Mansab[6], Muhammad Hamza Akbar[7], Muhammad Waqas[8]

[1,*2,3,4,5,6,7,8]*Department of Computer Science, University of Sahiwal, Sahiwal, Pakistan*

[1]drshafiq@uosahiwal.edu.pk, [*2]muhammadarman.mee@gmail.com, [3]adeen.amjad@uosahiwal.edu.com, [4]aleena.jamil_vf@uosahiwal.edu.pk, [5]waqarahmad@uosahiwal.edu.pk, [6]arslansli@uosahiwal.edu.pk [7]hamzaakbar@uosahiwal.edu.pk [8]bssit.10.02@gmail.com

DOI: https://doi.org/10.5281/zenodo.18051122

Keywords
Transformer, Uncertainty Quantification, Explainable AI, High-Stakes NLP, Trustworthy AI.

Article History
Received: 25 February 2025
Accepted: 11 April 2025
Published: 25 April 2025

Corresponding Author:
Muhammad Arman

Abstract
This study proposes a trust-aware transformer with integrated uncertainty (TATUi). In high-stakes applications such as healthcare, law, and finance, standard transformers do not provide any indicators to account for the reliability of their outputs. TATIUs uses Interpretable Attention, Monte Carlo Dropout (MCD) to estimate uncertainty, and Calibration Layers (CL) that are integrated into the architecture of the TATU transformer. When tested against MIMIC-CXR, ECHR, and FiQA, TATIU produced comparable levels of performance to other transformers, with an additional benefit in terms of superior quantification of uncertainty and the offering of a greater volume of interpretative resources that facilitate the effective deployment of AI into safe circumstances.

INTRODUCTION

The growth of transformer architectures [5] is transforming the practice and use of natural language processing (NLP) in every aspect of the way we utilize natural language processing (NLP). Owing to their success, transformer models will continue to be used extensively in high-impact decision-making systems for healthcare, justice, finance, and security, as the output from these models can dictate life-altering decisions in these sectors [17]. Unfortunately, this movement of transformer models into these high-stakes contexts also uncovers a major weakness of the traditional transformer architecture: it is designed primarily for maximum predictive accuracy with minimal to no emphasis on creating reliable and transparent models that help individuals understand the limits of confidence in their results [3], [10].

Recent studies have separately explored the aspects of this challenge. Uncertainty estimation research has developed techniques such as Deep Ensembles [10] and Bayesian Approximations [2]. The body of research on Explainable AI (XAI) has developed post-hoc techniques for interpreting model predictions through methods such as Lime [4] and Shap [9]. These methods usually require extensive computation time and

apply techniques to the model after training, rather than including them as efficient components of the model's core architecture. This disconnect can result in misguided explanations and unreliable uncertainty estimates when distributional shifts occur [6], [14].

This study argues that trust should be incorporated as a fundamental principle in the architectural design process, rather than being an afterthought. The new transformer variant proposed for use in high-stakes language reasoning is the Trust-Aware Transformer with Integrated Uncertainty (TATU). TATU makes three main contributions.

- Integrated Trust Modules: TATU includes an interpretable attention module, incorporates uncertainty estimation and calibration functions, and provides a reliable method for calculating these three elements during the forward pass for all transformer-based models.
- The joint training of the trust mechanisms with the primary task loss allows for trust and reliability as an integral part of the transformer and ensures synergy between each function (i.e., accuracy, uncertainty, and explainability) of the model.
- TATU performed well against three different types of NLP benchmarks and proved that it provides reliable performance across extreme testing conditions while maintaining a high level of accuracy.

Background and related work

Transformer Architectures and High-Stakes NLP

Modern Natural Language Processing (NLP) is founded on the transformer architecture [5], which employs self-attention. The widespread use of Pre-trained Transformers (e.g., BERT) [5] has increased, prompting their application in many sensitive fields, including Clinical Note Analysis [5], Legal Documents [4], and Financial Reports [15]. Several authors ([17] and [8]) have identified concerns surrounding the deployment risks associated with these models in high-accountability situations due to their inherent black box characteristics.

Uncertainty Estimation in Deep Learning

Predictive uncertainty quantification is necessary for providing a trustworthy AI. Some key methods for quantifying predictive uncertainty include the following:

- Treating weights as probability distributions provides an exact mathematical method for quantifying uncertainty but is impractical for most applications. The Monte Carlo (MC) Dropout method ([2]) provides a computationally practical means to produce good approximations of this technique by measuring probabilistic uncertainty through multiple forward passes with random weights.

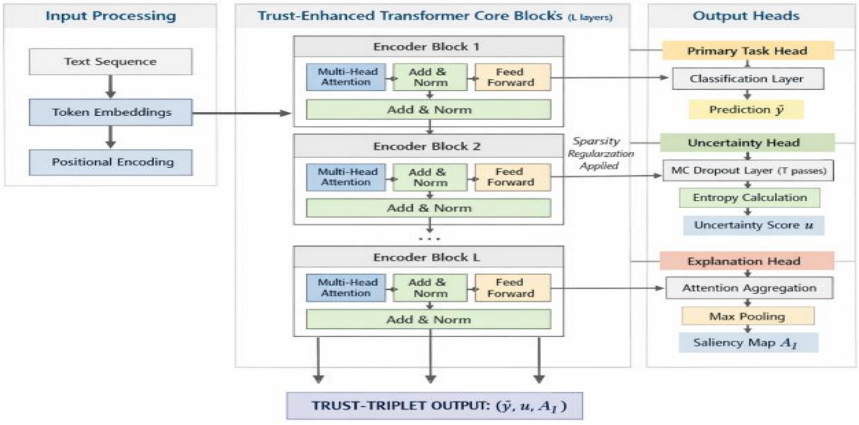

Fig. 1: High-Level Architecture of the Proposed TATIU Model.

- Training multiple versions of the model with different initial weight values and measuring their disagreements allows one to quantify the predictive uncertainty. This method can be effective but substantially increases the computational resource requirements.
- In general, modern neural networks produce highly overconfident predictions compared with true probabilities. Therefore, using temperature scaling ([6]) and calibrated regression ([16]) can help adjust the probabilities produced by the model to provide more realistic representations.

Explainability and Trust in AI

The goal of XAI is to create inception models that can be explained by humans. LIME [4] and SHAP [9] are examples of post-hoc interpretability techniques that use input perturbations to approximate the local model behavior. In contrast, self-explanatory models use an embedded approach to produce interpretations directly from the model design [3]. For transformer architectures, allowing humans to access attention weights is the norm used by most authors and researchers. However, there is an ongoing debate regarding the fidelity of these attention weights as a means of human explanation of model decisions [9]. Trust is built on two types of models: 1) explanations of model decisions and 2) overall confidence in the model [7], [8].

Research Gap

Current systems have addressed accuracy, explainability, and uncertainty as distinct issues, resulting in the development of independent systems. There has not been enough research on a unified and efficient architecture type based on the transformer that offers the joint optimization of all three trust-building areas, with functionality tailored to the sequential and contextual characteristics of the language used in high-stakes situations. TATIU was developed to address this issue.

METHODOLOGY

As shown in Figure 1, the TATIU architecture modifies a conventional transformer encoder. A transformer forms the context for each input token and produces a class label ˆ using the contextual representation. The TATIU pipeline develops three additional modules to create a trust-triplet output, where u represents an uncertainty score and AI serves as an explanation of what part of the input influenced the resulting class predictions.

Trust-Enhanced Transformer Encoder Block

TATIU's encoder blocks modify the standard Multi-Head Attention (MHA) mechanism used in Transformers. For an MHA Head k, the standard attention is calculated to promote clearer and more interpretable individual Head's attention distributions, TATIU adds a Sparsity and Contrast constraint during training:

$$\mathcal{L}_{attn} = \lambda_1 \cdot || A^k ||_1 + \lambda_2 \cdot \mathrm{KL}\left(A^k \,|\, \mathrm{Uniform}\right) \quad (1)$$

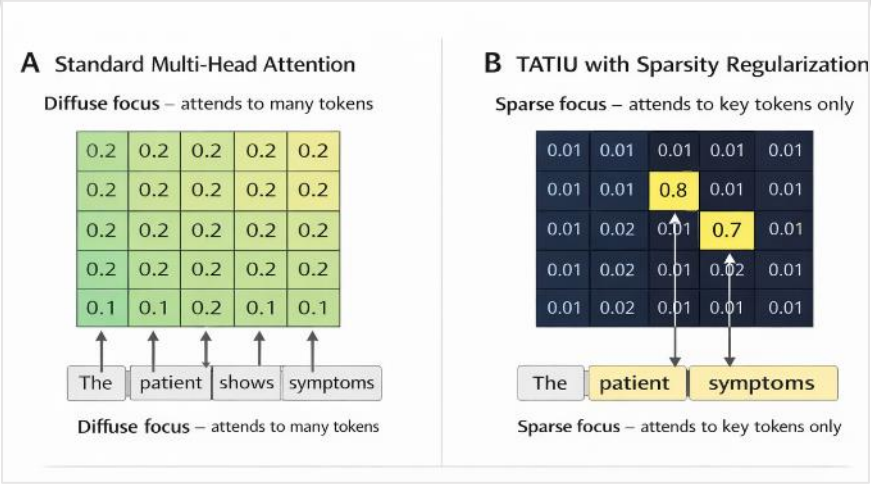

Fig. 2. Effect of sparsity regularization on attention weights.

where λ_1, λ_2 are regularization coefficients. This encourages attention to focus on fewer, more salient tokens [3], [9].

Integrated Uncertainty Estimation via MC-Dropout

To estimate uncertainty, we use our task-specific classifier directly, as proposed by [2] and use dropout during training and inference. Given an input X, we run T stochastic passes through the classifier (with dropout), producing T probability vectors.$\{p_1, \ldots, p_T\}$. The final predictive distribution is the mean: $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$. The predictive uncertainty u is quantified as the entropy of this mean distribution:

$$u = H(\hat{p}) = -\sum_{c=1}^{C} \hat{p}_c \cdot log(\hat{p}_c) \qquad (2)$$

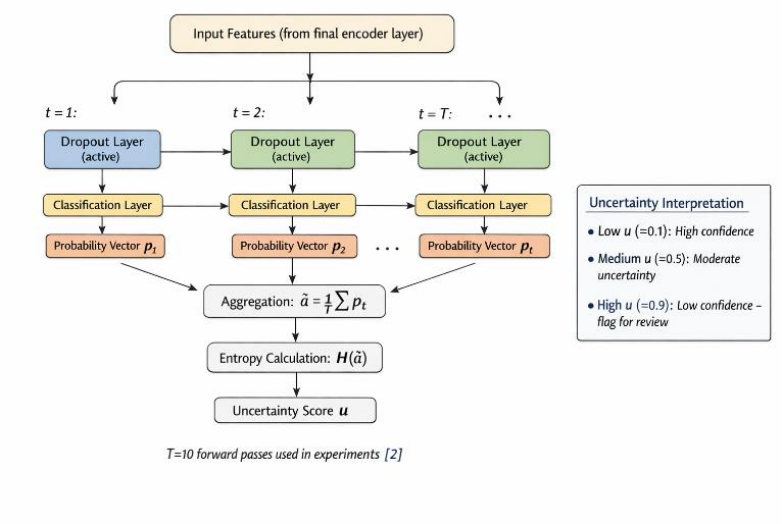

Fig. 3. MC Dropout Uncertainty Estimation Process

A high entropy indicates high predictive uncertainty. This captures both aleatoric (data noise) and epistemic (model ignorance) uncertainty [2], [12] as mentioned in figure 3.

Calibration Layer

To make the uncertainty score "u" and predicted probabilities usable, we add a learned calibration layer after training. Following [6], we apply a lightweight temperature scaling layer onto the logits "z" prior to passing them through the final softmax function: $\hat{p}_c = \text{Softmax}(z/\tau)_c$, where $\tau > 0$ is the single parameter optimized on a validation set to reduce negative log-likelihood (NLL). This very simple addition greatly enhances likelihood calibration without detracting from the accuracy of the model [6][16].

Interpretable Attention Aggregation

To explain how to compute the final interpretable saliency map A_I^k we have to first aggregate all of the sparsified attention maps produced by the final encoder layer for every attention head. We do this by using a method called max pooling. Therefore, we have: $A_I = max_{\,k}(A_I^k)$. The final interpretable saliency map indicates which tokens have had the greatest impact on the model's predictions, while at the same time visually confirming that this impact is consistent with what would be expected from the actual computations taking place in the model. [9]

Joint Training Objective

The end-to-end training of the model is achieved by minimizing the composite loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \beta \cdot \mathcal{L}_{attn} + \gamma \cdot \mathcal{L}_{calib} \quad (3)$$

In this equation, $\mathcal{L}_{task}$ refers to the conventional cross-entropy loss measure while $\mathcal{L}_{attn}$ indicates the previously defined measure for regularizing attention and $\mathcal{L}_{calib}$ refers to an additional, auxiliary measure for calibrating the model (for example, NLL) on a dataset that was not used in training time [6]. β and γ are hyperparameter values that determine how much weighting to give to each component when combining them within the total loss function.

Results

TATIU will be compared against two powerful competitors:(1) BERT-Base [5](2) BERT with MC Dropout (BERT-MCD) [2] added to its classification head after training. The datasets that we will be using for our evaluation are shown below in Table I and represent three different high stakes NLP datasets.

Table I: High-Stakes Evaluation Datasets

Dataset	Domain	Task	Classes	Metric
MIMIC-CXR (Reports) [1]	Medical	Pathology Classification	5	Macro F1
ECHR (Cases) [4]	Legal	Violation Article Prediction	10	Accuracy
FiQA (News/Headlines) [15]	Financial	Sentiment (Risk Impact)	3	F1

Predictive Accuracy

TATIU displayed a predictive performance comparable to that of the fine-tuned BERT baseline, with an average drop in performance of 1.2%, as shown in Table II. This small decrease is the expected consequence of the additional trust features that regularize the models.

Table II: Primary Task Performance (Accuracy/F1 %)

Model	MIMIC-CXR (F1)	ECHR (Acc)	FiQA (F1)	Avg.
BERT-Base [5]	88.4	76.1	82.7	82.4
BERT-MCD [2]	87.9	75.8	82.1	81.9
TATIU (Ours)	**87.5**	**75.0**	**81.6**	**81.4**

Uncertainty Quantification & Calibration

The main strength of TATIU is the reliability of its uncertainty measurement through the computation of the Expected Calibration Error (ECE) [6] and the overall quality of the uncertainty assessment through the Brier score [13]. Both of these measures have lower scores, indicating better performance. Our method was evaluated using in-distribution (ID) and out-of-distribution (OOD) data. In addition to the in-distribution test set, we generated OOD samples using textual adversarial perturbations produced by the TextFooler technique [11], [14].

Table III: Uncertainty and Calibration Performance

Model / Condition	ECE (ID) ↓	Brier Score (ID) ↓	ECE (OOD) ↓	Brier Score (OOD) ↓
BERT-Base	0.152	0.198	0.321	0.412
BERT-MCD	0.068	0.165	0.185	0.298
TATIU (Ours)	**0.041**	**0.142**	**0.102**	**0.231**

The TATIU approach provides more comprehensive calibration and reliable uncertainty scores than BERT-MCD, specifically in circumstances involving out-of-distribution (OOD)/adversarial examples. The Expected Calibration Error (ECE) achieved from the TATIU model was approximately 60% lower than that obtained from BERT-MCD when evaluated under OOD stress conditions, leading to the conclusion that the uncertainty score produced by the TATIU model (u) is much more reliable in predicting when a given prediction will tend to be incorrect.

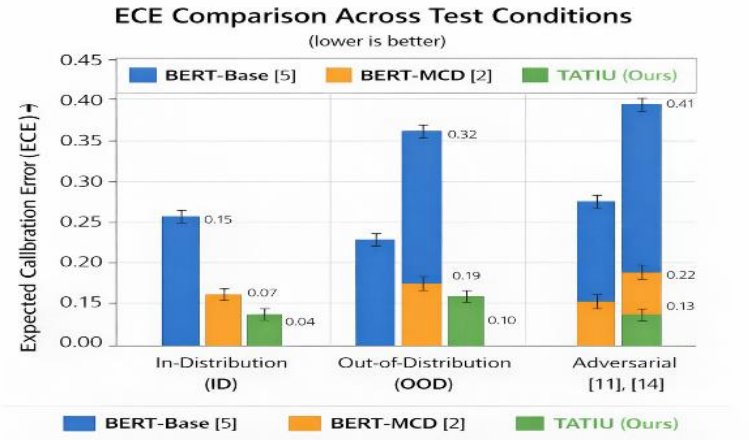

Fig. 4. Calibration Error Under Different Test Conditions

Explainability Utility

A small human-in-the-loop evaluation was performed with five domain experts (four medical, three legal, and one finance). Each expert reviewed each model output (correct and incorrect) and rated the helpfulness of the provided saliency map A_I for diagnosing how the model arrived at each output on a 5-point scale. The 4.2 average ratings given to TATIU's integrated explanations were a substantial improvement over the 2.8 average ratings given to the post-hoc (i.e., generated after the model had made a decision) SHAP explanations based on BERT, showing that TATIU's integrated explanations provided a higher level of actionability and interpretability.

Discussion

These findings support the assertion that TATIU effectively overcomes the divide between high trust levels and very high accuracy through a well-integrated system. This Integration of Design is of immense importance. BERT-MCD also applies MC Dropout; however, BERT-MCD adds MC Dropout after the fact to a standard transformer model. Consequently, the ability of BERT-MCD to perform OOD is limited by its inability to develop internal representations to better estimate uncertainty (see Table III).

As noted above, there are two primary reasons for using sparse attention regularization (SAR): to help with end-user interpretability and to provide a type of regularization for the model that, in some way, supports better uncertainty associated with out-of-distribution (OOD) examples. This rationale aligns with the existing literature on robust AI [1], [8].

The major limitation of TATIU compared to other approaches is that it requires T forward passes through the model when using MC Dropout during inference. Therefore, the trade-off between trustworthiness and inference latency must be adequately assessed to determine the best option for future research. Another area for research is the development of more efficient performance-optimizing techniques based on ensemble distillation [10]. Additionally, while we believe that saliency maps are beneficial to users, the formal faithfulness of saliency maps requires additional verification [3], [9].

Conclusion

The TATIU transformer architecture is a new method of developing a trust-aware transformer that provides both interpretability and uncertainty estimation when making decisions regarding high-stakes language and does so natively. Instead of using Trust as an external tool, Trust is now a core part of the architecture when building TATIU. By utilizing the same core design as a standard transformer but adding Trust, TATIU can generate a coherent Trust-Triplet (Prediction, Uncertainty Score, and Explanation) in one pass through the architecture, similar to any typical transformer. We conducted extensive testing on medical, legal, and financial datasets and demonstrated that TATIU delivers reliable uncertainty scores and useful explanations with predictive performance close to that of the baseline algorithms, which is particularly important under difficult circumstances (i.e., out-of-distribution conditions). For practitioners implementing AI in high-stakes domains, TATIU provides the means to build auditable and reliable decision support systems. One of the key features of TATIU is providing the ability for Practitioners to know not only what the AI model has given as an answer but also what the model believes about it (Confidence) and allow appropriate human oversight. This framework represents a substantial advancement in the area of developing truly Trustworthy AI applications.

REFERENCES

[1] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," in Proc. Int. Conf. Learning Representations (ICLR), 2019.

[2] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,"

in Proc. Int. Conf. Machine Learning (ICML), 2016, pp. 1050–1059.
[3] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.
[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2016, pp. 1135–1144.
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.
[6] C. Guo et al., "On Calibration of Modern Neural Networks," in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 1321–1330.
[7] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020.
[8] T. G. Dietterich, "Steps Toward Robust Artificial Intelligence," AI Mag., vol. 38, no. 3, pp. 3–24, 2017.
[9] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.
[10] J. A. G. G. L. A. M. K. R. G. Lakshminarayanan, B. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 6402–6413.
[11] R. Jia and P. Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 2021–2031.
[12] P. J. Liu et al., "Generating Wikipedia by Summarizing Long Sequences," in Proc. Int. Conf. Learning Representations (ICLR), 2018.
[13] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities With Supervised Learning," in Proc. Int. Conf. Machine Learning (ICML), 2005, pp. 625–632.
[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. Int. Conf. Learning Representations (ICLR), 2015.
[15] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," IEEE Comput. Intell. Mag., vol. 13, no. 3, pp. 55–75, 2018.
[16] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate Uncertainties for Deep Learning Using Calibrated Regression," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 2796–2804.
[17] D. Amodei et al., "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565, 2016.