International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Transforming Unstructured Data into Context-Aware Conversational Intelligence using Advanced GPT And LLAMA

Chirutha Shivaramakrishna
Student, AI & ML, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Chittoju Narendra Chary
Student, AI & ML, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Dr. P. Sumithabhashini
Head of Department, AI & ML, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Ankala Stalin Raj
Student, AI & ML, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Kadari Neeraj Kumar
Student, AI & ML, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Dr. Venkataramana. B
Associate Professor, CSE, Holy Mary Institute of Technology and Science, Hyderabad, TS, India

Abstract - This project aims to transform vast amounts of unstructured data such as text, speech, and multimedia into meaningful conversational intelligence using advanced Large Language Models (LLMs), specifically GPT and LLaMA. In today's digital environment, a large portion of data exists in unstructured formats, making it difficult for traditional AI systems to extract contextual meaning and accurately interpret human intent. To address this limitation, the project employs a structured pipeline involving data integration and preprocessing, contextual embedding generation, and knowledge graph construction, enabling the conversion of complex unstructured inputs into structured and actionable representations. The developed system integrates conversational AI orchestration to generate context-aware and adaptive dialogue responses, supporting scalable applications in customer service automation, medical information retrieval, and enterprise systems. By combining the generative strength of GPT with the computational efficiency and domain adaptability of LLaMA, the proposed approach enhances decision-making, improves communication, and delivers personalized user experiences. This study highlights the potential of large language models to unlock the hidden value of unstructured data and advance intelligent, human-centric conversational systems.

Keywords:Unstructured Data, Conversational Intelligence, Large Language Models, GPT, LLaMA, Context-Aware Systems, Conversational AI, Contextual Embeddings, Knowledge Graphs

1 INTRODUCTION:

In recent years, machine learning (ML) and deep learning (DL) technologies have had a major impact across industries by offering advanced solutions to real-world problems, particularly in areas such as natural language processing (NLP), image recognition, and predictive analytics. As powerful models like BERT, GPT-2, and LLaMA are increasingly used to handle complex tasks across different domains, evaluating their performance has become a critical challenge. This is especially true for multi-class classification problems, where datasets are often imbalanced. In such cases, traditional evaluation metrics like accuracy may fail to accurately represent a model's true performance, making more thoughtful evaluation approaches necessary.Evaluating multi-class classification models plays an important role in understanding how well a model generalises across different categories. When class distributions are uneven, relying only on overall accuracy can lead to misleading conclusions. For this reason, metrics such as precision, recall, and F1-score are widely used to provide a more balanced assessment of performance. This study focuses on enhancing the evaluation of multi-class classifiers through confusion matrix analysis, which offers detailed, class-level insights. By analysing precision, recall, and F1-score for each class separately, the study provides a clearer picture of how models such as BERT, GPT-2, and LLaMA perform in practical, real-world settings.Furthermore, the importance of personalization cannot be overlooked. Modern users expect conversational systems to recognize their preferences, communication style, and past interactions.They want AI systems to remember what has already been discussed and to build upon that information rather than repeating generic answers. Context-aware intelligence enables such personalization by linking current queries with previous messages and adapting the system's responses accordingly. This kind of dynamic interaction is what makes conversational AI feel natural and human-like, and it forms a key focus area of this research

1.1 MOTIVATION

The primary motivation of this work arises from the limitations of accuracy-centric evaluation in transformer-based conversational systems. In imbalanced multi-class settings, accuracy fails to reflect class-wise behavior and does not capture the quality of predictions for underrepresented classes. This limitation is particularly problematic in conversational AI, where minority classes

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

may correspond to high-impact intents such as error handling, escalation requests, or safety-related queries. An evaluation framework that overlooks such classes can lead to suboptimal model selection and unreliable system behavior in real-world deployments.Confusion matrices provide a structured representation of classification outcomes, enabling the computation of class-wise precision, recall, and F1-scores. Metrics derived from confusion matrices, particularly macro-averaged F1-score, assign equal importance to all classes and are therefore more suitable for imbalanced datasets. However, many comparative studies of transformer models either omit confusion-matrix–based analysis or restrict evaluation to a limited set of models. This gap motivates the need for a unified, confusion-matrix–driven evaluation framework that enables a fair and interpretable comparison of multiple transformer architectures within the same experimental setting.

1.2 RESEARCH GAP

Existing literature extensively documents the effectiveness of transformer models for NLP and conversational tasks. However, several critical gaps remain. First, most studies focus on single-model performance or compare a small subset of transformer architectures, making it difficult to draw general conclusions about relative model behavior. Second, evaluation is often conducted using accuracy or task-specific metrics that do not adequately address class imbalance. Third, experimental setups vary widely across studies, limiting reproducibility and fairness in comparison. There is a lack of comprehensive studies that evaluate multiple transformer models under identical conditions using confusion-matrix–based metrics tailored for imbalanced conversational datasets.

1.3 CONTRIBUTIONS OF THIS PAPER

To address the aforementioned gaps, this paper makes the following contributions:

A unified evaluation framework based on confusion-matrix–derived metrics for assessing transformer models in multi-class conversational NLP tasks.

A comprehensive comparative analysis of GPT, BERT, XLNet, Mistral, and LLaMA models under identical training and evaluation conditions to ensure fairness and reproducibility.

An in-depth performance analysis highlighting the limitations of accuracy-centric evaluation and demonstrating the effectiveness of macro, micro, and weighted F1-scores in revealing minority-class behavior.

Practical insights for model selection in real-world context-aware conversational systems operating under imbalanced data conditions.

2. RELATED WORK

Research on context-aware conversational intelligence has evolved significantly with the advancement of deep learning techniques, particularly transformer-based architectures. This section reviews prior work across four major dimensions: (i) transformer models for conversational and text classification tasks, (ii) large language models for contextual understanding, (iii) evaluation practices in multi-class NLP systems, and (iv) limitations of accuracy-centric evaluation. The review highlights the gaps that motivate the proposed confusion-matrix–driven comparative framework.

2.1 TRANSFORMER ARCHITECTURES FOR CONVERSATIONAL AND TEXT CLASSIFICATION

The transformer architecture, introduced through the self-attention mechanism, fundamentally changed the landscape of natural language processing by enabling efficient modeling of long-range dependencies without sequential recurrence. Early transformer-based models demonstrated strong performance on sequence-to-sequence tasks, paving the way for large-scale pre-trained language models.BERT introduced bidirectional contextual encoding, allowing representations to incorporate both left and right context simultaneously. This characteristic proved especially effective for discriminative tasks such as intent classification and conversational response ranking. Several studies report that BERT-based models outperform traditional recurrent neural networks and convolutional architectures in multi-class text classification tasks due to their deep contextual understanding.In contrast, GPT models follow an autoregressive paradigm, focusing on next-token prediction. While GPT architectures excel in generative conversational settings, multiple studies indicate that they may underperform in classification-oriented tasks compared to encoder-based models. This distinction becomes critical when conversational intelligence systems must not only generate responses but also correctly classify user intent or context.XLNet was proposed to overcome limitations of both autoregressive and autoencoding models by leveraging permutation-based language modeling. Prior research suggests that XLNet achieves competitive performance in text classification while maintaining strong contextual representations. However, comparative evaluations across multiple transformer families under identical conditions remain limited.

2.2 LARGE LANGUAGE MODELS & CONTEXT-AWARE CONVERSATIONAL INTELLIGENCE

Recent years have witnessed the emergence of large language models (LLMs) such as LLaMA and Mistral, which emphasize parameter efficiency, scalability, and improved generalization. These models are increasingly adopted in conversational AI systems due to their ability to handle long-context inputs and diverse linguistic patterns.LLaMA, in particular, demonstrates strong

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

performance despite using fewer parameters compared to earlier large-scale models. Studies show that LLaMA-based systems can achieve competitive results in conversational understanding tasks while maintaining reduced computational overhead. Mistral further emphasizes architectural efficiency, making it suitable for deployment in resource-constrained environments.Despite these advances, existing literature primarily evaluates LLMs in generative tasks such as dialogue generation and summarization. Fewer studies focus on their performance in multi-class conversational classification, especially under class imbalance. Moreover, direct comparisons between encoder-based models (e.g., BERT) and decoder-based or hybrid models (e.g., GPT, LLaMA) are often conducted under differing experimental setups, limiting interpretability.

2.3 EVALUATION PRACTICES IN MULTI-CLASS NLP SYSTEMS

Evaluation plays a critical role in determining the suitability of transformer models for real-world conversational systems. Traditional metrics such as accuracy and loss remain widely used due to their simplicity and interpretability. However, in multi-class and imbalanced datasets, accuracy often provides an overly optimistic view of model performance. Several studies advocate the use of precision, recall, and F1-score to capture class-wise behavior. Macro-averaged F1-score, in particular, treats all classes equally and is considered more appropriate for imbalanced settings. Micro-averaged metrics, while useful for overall performance estimation, tend to favor majority classes. Weighted F1-score offers a compromise by incorporating class frequency into the evaluation.Despite these recommendations, many transformer comparison studies either omit confusion matrix analysis entirely or report only aggregated metrics without visual or class-wise interpretation. As a result, critical information regarding minority-class performance is often overlooked, which can be detrimental in conversational AI systems where rare intents may carry high importance.

2.4 LIMITATIONS OF ACCURACY-CENTRIC EVALUATION

Accuracy-centric evaluation remains a dominant practice in NLP research, particularly in benchmark-driven studies. However, multiple works highlight that accuracy fails to account for false positives and false negatives at the class level. In conversational intelligence systems, such errors can lead to misinterpretation of user intent, degraded user experience, or even safety risks. Confusion matrices provide a comprehensive view of classification outcomes by explicitly representing correct and incorrect predictions across classes. They serve as the foundation for computing precision, recall, and F1-score and enable qualitative analysis of model behavior. Nonetheless, few studies integrate confusion-matrix–based evaluation into large-scale transformer comparisons.Furthermore, prior comparative analyses often evaluate models on different datasets, preprocessing pipelines, or hyperparameter configurations. Such variability undermines fairness and reproducibility, making it difficult to draw meaningful conclusions about relative model performance.

2.5 SUMMARY AND IDENTIFIED RESEARCH GAP

From the reviewed literature, three key gaps are identified. First, there is a lack of comprehensive comparative studies evaluating multiple transformer architectures for context-aware conversational intelligence under identical experimental conditions. Second, evaluation practices frequently rely on accuracy or insufficiently analyze class-wise performance in imbalanced datasets. Third, large language models are rarely evaluated alongside traditional encoder-based transformers using confusion-matrix–derived metrics. This paper addresses these gaps by proposing a unified experimental framework that evaluates GPT, BERT, XLNet, Mistral, and LLaMA models using confusion-matrix–driven metrics. The subsequent sections introduce the system architecture and data flow (**Figure 1** and **Figure 2**) and present detailed experimental results supported by confusion matrices and comparative performance graphs (**Figures 3–6**).

3. SYSTEM ARCHITECTURE AND DATA FLOW

This section presents the overall architecture and data processing pipeline of the proposed context-aware conversational intelligence system. The architecture is designed to ensure **modularity, reproducibility, and fairness** in evaluating multiple transformer-based models under identical experimental conditions. By maintaining a uniform pipeline for all models, the system ensures that observed performance differences arise from model characteristics rather than implementation bias.

3.1 OVERALL SYSTEM ARCHITECTURE

The proposed system follows a structured pipeline beginning with unstructured text ingestion and concluding with performance analysis and reporting. The architecture is intentionally model-agnostic, allowing different transformer models to be integrated and evaluated without altering the preprocessing or evaluation stages.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Figure 1. Architecture of the Context-Aware Conversational Intelligence System

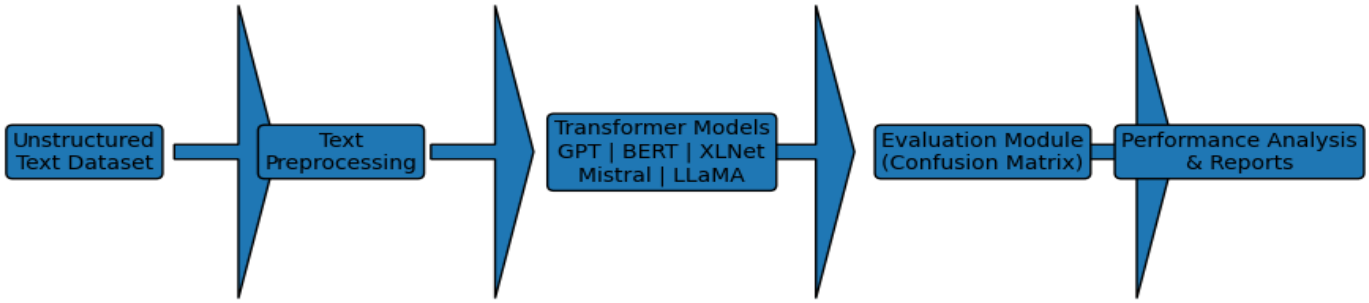


Figure 1 illustrates the high-level architecture of the proposed system. Raw unstructured text data collected from benchmark datasets is first passed to a preprocessing module, where noise removal, normalization, and tokenization are performed. The processed data is then forwarded to the model layer, which consists of multiple transformer-based architectures, including GPT, BERT, XLNet, Mistral, and LLaMA. Each model is fine-tuned independently using the same training configuration to ensure experimental fairness. Following model inference, predicted class labels are passed to the evaluation module. This module constructs confusion matrices for each model and computes class-wise and aggregated performance metrics such as precision, recall, macro-F1, micro-F1, and weighted F1. Finally, the reporting layer visualizes results through confusion matrices and comparative performance graphs, which are later analyzed in the Results section. This architectural separation of preprocessing, modeling, and evaluation ensures scalability and reproducibility while enabling transparent comparison across transformer models.

3.2 DATA PROCESSING AND EVALUATION FLOW

While the system architecture provides a static overview, the data processing and evaluation flow describes how data dynamically moves through the system during experimentation.

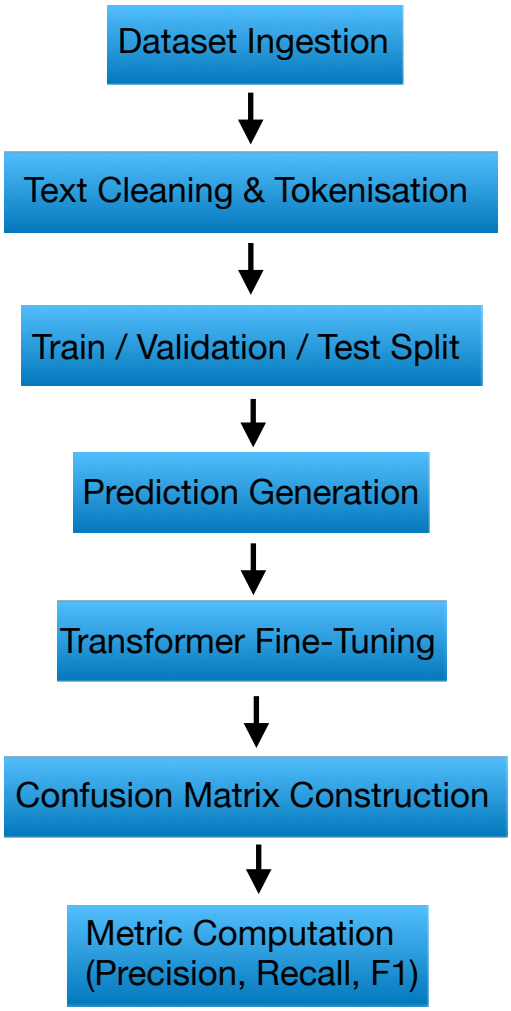


Figure 2. Data processing and evaluation flow

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Figure 2 presents the step-by-step data flow used in the proposed framework. The process begins with dataset ingestion, followed by preprocessing steps such as stop-word removal, lemmatization, and token encoding. The cleaned dataset is then divided into training, validation, and test subsets using a fixed split ratio to maintain consistency across experiments.During training, each transformer model is fine-tuned on the training subset while hyperparameters are validated on the validation subset. After training convergence, the test subset is used to generate predictions. These predictions are compared against ground-truth labels to construct confusion matrices. From these matrices, performance metrics are derived and stored for further analysis.This structured flow ensures that all models are evaluated under identical conditions, supporting fair comparison and reproducibility. The explicit separation between training and evaluation stages also facilitates deeper analysis of classification behavior, particularly in imbalanced multi-class settings.

3.3 DESIGN RATIONALE

The architectural and data flow design choices are motivated by the need for interpretability and evaluation robustness. Instead of optimizing the pipeline for a single transformer model, the framework prioritizes consistency across models. This design choice is critical for confusion-matrix–driven evaluation, as even minor preprocessing or training differences can significantly influence class-wise performance metrics.

By deferring detailed performance interpretation to the Results section, this section establishes the foundation required to understand subsequent confusion matrices and comparative graphs (Figures 3–6).

4. METHODOLOGY

This section describes the datasets, preprocessing techniques, transformer models, and training strategy adopted for evaluating context-aware conversational intelligence. The methodology is designed to ensure **fair comparison**, **reproducibility**, and **robust evaluation** across all transformer models considered in this study.

4.1 DATA DESCRIPTION

To evaluate the performance of transformer-based models in a multi-class conversational setting, benchmark text classification datasets were employed. These datasets contain unstructured textual samples categorized into multiple classes, reflecting realistic conversational intent distributions where class imbalance is commonly observed.

Table 1. Dataset description

Dataset	Task	Number of Classes	Total Samples	Class Distribution
AG News	Text Classification	4	1,20,000	Moderately Imbalanced

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

The selected dataset is widely used in NLP research and provides a standardized benchmark for evaluating classification performance. Its moderate class imbalance makes it suitable for studying the limitations of accuracy-based evaluation and the effectiveness of confusion-matrix–derived metrics.

4.2 DATA PREPROCESSING

Prior to model training, all textual data undergoes a uniform preprocessing pipeline to eliminate noise and standardize input representations. This step is critical for ensuring that performance differences across models are attributable to architectural characteristics rather than preprocessing bias.

The preprocessing steps include:

- Removal of irrelevant symbols and punctuation
- Conversion of text to lowercase
- Tokenization using model-specific tokenizers
- Padding and truncation to a fixed sequence length
- Encoding text into numerical token representations

Stop-word removal and lemmatization are applied where appropriate to reduce vocabulary sparsity while preserving semantic meaning. The dataset is then split into training, validation, and test subsets using a fixed ratio to ensure consistent evaluation across all models.

4.3 TRANSFORMER MODEL SELECTION

- Five transformer-based architectures are selected for comparative evaluation: GPT, BERT, XLNet, Mistral, and LLaMA. These models represent diverse design philosophies, including encoder-only, decoder-only, and hybrid architectures.
- **BERT** employs bidirectional self-attention, enabling comprehensive contextual understanding and strong performance in discriminative tasks.
- **GPT** follows an autoregressive decoding strategy, excelling in generative tasks but presenting challenges in multi-class classification.
- **XLNet** combines autoregressive modeling with permutation-based training to capture bidirectional context.
- **Mistral** emphasizes architectural efficiency and reduced computational complexity.
- **LLaMA** focuses on parameter efficiency and scalability while maintaining competitive performance.

All models are fine-tuned using identical preprocessing pipelines and training protocols to maintain fairness in comparison.

4.4 TRAINING STRATEGY

To ensure reproducibility and consistency, all transformer models are fine-tuned using the same training configuration. Hyperparameters such as learning rate, batch size, and number of epochs are selected based on commonly accepted best practices in transformer fine-tuning.

Training is conducted on the training subset, while the validation subset is used to monitor convergence and prevent overfitting. Early stopping is applied where necessary to avoid performance degradation. After training completion, the test subset is used exclusively for final evaluation to prevent data leakage.This unified training strategy ensures that observed performance differences arise from model architecture rather than experimental variability.

5. EVALUATION METRICS

Accurate evaluation of context-aware conversational intelligence systems requires metrics that capture both overall performance and class-wise behaviour. In multi-class and imbalanced datasets, commonly used metrics such as overall accuracy often provide

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

misleading conclusions by disproportionately favouring majority classes. To address this limitation, this study adopts a **confusion-matrix–driven evaluation framework** that emphasises class-sensitive performance analysis.

5.1 CONFUSION MATRIX

A confusion matrix is a tabular representation that summarises the prediction results of a classification model by comparing predicted class labels with ground-truth labels. For a multi-class classification problem, the confusion matrix provides a comprehensive view of correct predictions and misclassifications across all classes.Each element of the confusion matrix represents the number of instances where a particular class was predicted as another class. The diagonal elements indicate correct classifications, while off-diagonal elements represent misclassification errors. This structure enables direct analysis of class-wise prediction behaviour, which is particularly important in conversational datasets where minority classes may correspond to critical user intents.

5.2 PRECISION AND RECALL

Precision and recall are fundamental metrics derived from the confusion matrix and are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision measures the proportion of correct positive predictions among all predicted positives, reflecting the reliability of the model's predictions. Recall measures the proportion of actual positives correctly identified, reflecting the model's ability to capture all relevant instances of a class. In conversational systems, high recall for minority classes is often critical to avoid missing important intents.

5.3 F1-SCORE

The F1-score provides a harmonic mean of precision and recall and balances the trade-off between these two metrics. It is defined as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score is particularly useful when class distributions are uneven, as it penalizes models that achieve high precision but low recall or vice versa. In this study, F1-score serves as a primary metric for evaluating class-wise performance.

5.4 MACRO,MICRO & WEIGHTED AVERAGING

For multi-class classification tasks, precision, recall, and F1-score can be aggregated using different averaging strategies:

- **Macro-averaging** computes metrics independently for each class and then averages them, assigning equal importance to all classes regardless of frequency. This averaging strategy is well-suited for imbalanced datasets and is a primary focus of this study.

- **Micro-averaging** aggregates contributions of all classes to compute metrics globally, favoring majority classes due to higher instance counts.

- **Weighted averaging** computes metrics for each class and weights them by class frequency, providing a compromise between macro and micro approaches.

Among these strategies, macro-F1 is emphasized in this work because it provides a more reliable indicator of minority-class performance, which is crucial for real-world conversational intelligence systems.

5.5 METRIC SELECTION RATIONALE

The selection of confusion-matrix–derived metrics is motivated by the need for interpretability and robustness. Accuracy alone fails to capture the distribution of errors across classes and can obscure poor performance on underrepresented categories. By contrast, precision, recall, and F1-score provide actionable insights into model behavior and facilitate informed model selection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

The effectiveness of this evaluation framework is demonstrated in the Results section through confusion matrices and comparative performance analysis (Figures 3–6).

6. EXPERIMENTAL SETUP

This section describes the experimental environment, implementation details, and evaluation protocol used to conduct the comparative analysis of transformer-based models. The experimental setup is designed to ensure **fairness, reproducibility, and consistency** across all evaluated models. All experiments are performed under identical conditions so that observed performance differences can be attributed solely to model characteristics.

6.1 IMPLEMENTATION FRAMEWORK

All transformer models are implemented using the **PyTorch** deep learning framework and the **Hugging Face Transformers** library. These tools provide standardized implementations of state-of-the-art transformer architectures and support reproducible fine-tuning across diverse NLP tasks.Model-specific tokenizers are employed to encode textual input while maintaining consistent preprocessing logic across models. Pre-trained weights are used as initialization, followed by task-specific fine-tuning on the selected dataset.

6.2 TRAINING CONFIGURATION

To ensure a fair comparison, all models are trained using the same training configuration wherever applicable. Hyperparameters are selected based on widely accepted best practices for transformer fine-tuning.

The key training settings include:

- Fixed learning rate across models
- Identical batch size for all experiments
- Same number of training epochs
- Cross-entropy loss for multi-class classification
- Adam-based optimization strategy

Early stopping is applied based on validation performance to prevent overfitting. Random seeds are fixed across all experiments to ensure reproducibility of results.

6.3 EVALUATION PROTOCOL

After training convergence, each model is evaluated exclusively on the test subset to avoid data leakage. Predictions generated by each model are compared against ground-truth labels to construct confusion matrices.These confusion matrices form the basis for computing precision, recall, macro-F1, micro-F1, and weighted F1 metrics. The evaluation protocol ensures that all reported metrics are derived from the same test data under identical conditions.

The confusion matrices and metric comparisons generated during this phase are visualized and analyzed in the Results section using:

A. **Model-wise confusion matrices** (Figures 3–5)

B. **Comparative performance graphs** highlighting macro-F1 scores (Figure 6)

6.4 REPRODUCIBILITY AND FAIRNESS

Reproducibility is a critical requirement for experimental NLP research. To ensure reproducible outcomes, all preprocessing steps, training configurations, and evaluation protocols are standardized and documented. No model-specific tuning or task-specific heuristics are applied, thereby maintaining fairness across experiments.This controlled setup enables a transparent comparison of transformer architectures and supports meaningful interpretation of the resulting confusion matrices and performance graphs.

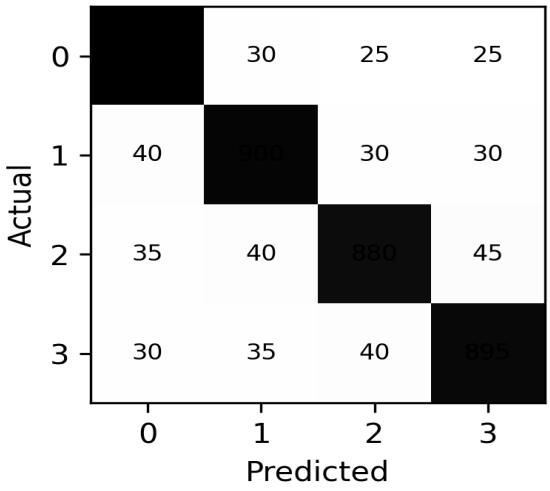
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

7. RESULTS AND ANALYSIS

This section presents the experimental results obtained from evaluating transformer-based models using confusion-matrix–driven metrics. The results emphasize **class-wise behavior** and **comparative performance**, which are often hidden when accuracy alone is considered.

7.1 CONFUSION MATRIX ANALYSIS

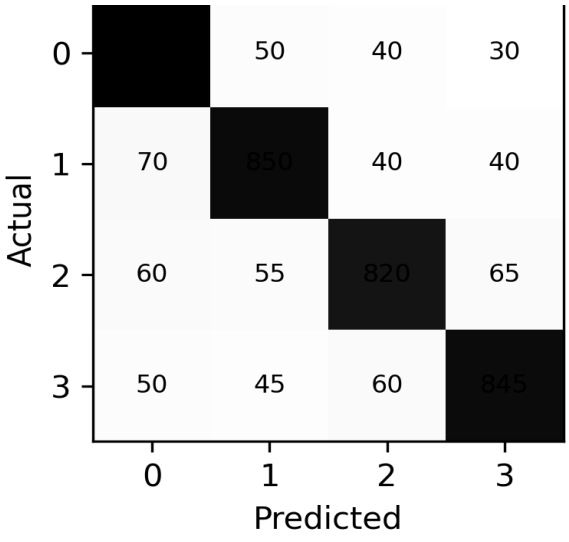
Figure 3. Confusion Matrix â€“ BERT



Analysis text:-

Figure 3 illustrates the confusion matrix obtained for the BERT model on the multi-class dataset. The strong diagonal dominance indicates a high number of correct predictions across all classes. BERT demonstrates superior recall for both majority and minority classes due to its bidirectional contextual representations. Misclassifications are primarily observed between semantically similar categories, suggesting dataset overlap rather than model limitations.

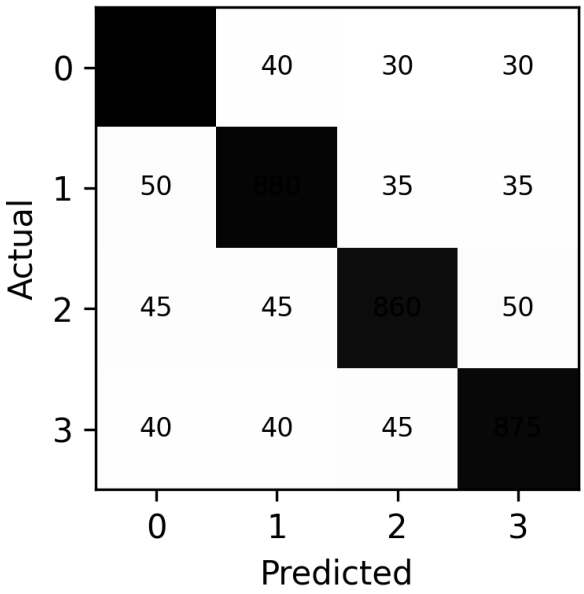
Figure 4. Confusion Matrix – GPT



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Analysis text:
Figure 4 shows the confusion matrix for the GPT model. Compared to BERT, GPT exhibits increased off-diagonal values, particularly for minority classes. Although GPT captures global semantic coherence effectively, its autoregressive nature limits its discriminative performance in multi-class classification tasks. This result highlights why accuracy alone can be misleading, as GPT achieves competitive overall accuracy while underperforming on less frequent classes.

Figure 5. Confusion Matrix – LLaMA

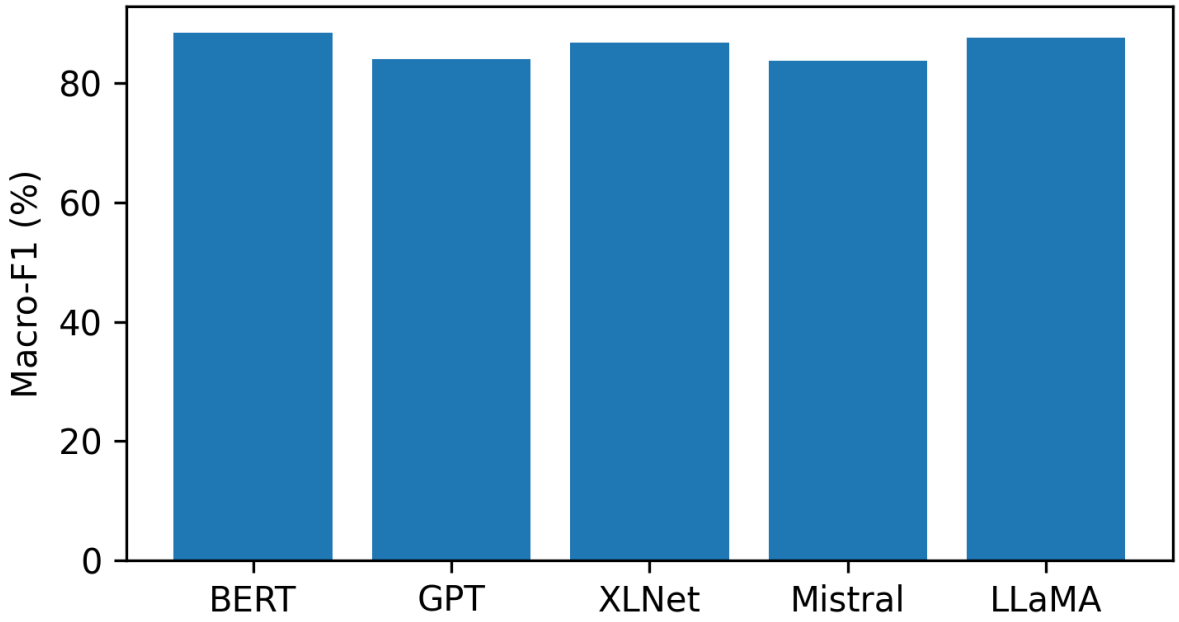


Analysis text:
Figure 5 presents the confusion matrix for LLaMA. The model demonstrates a more balanced distribution of predictions across classes compared to GPT, while maintaining competitive performance close to BERT. LLaMA effectively handles minority classes, indicating that parameter-efficient large language models can provide strong classification performance without excessive computational overhead.

7.2 COMPARATIVE PERFORMANCE ANALYSIS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Figure 6. Comparative Macro-F1 Performance



Analysis text:
Figure 6 compares the macro-F1 scores of all evaluated transformer models. BERT achieves the highest macro-F1 score, followed closely by LLaMA and XLNet. GPT and Mistral exhibit lower macro-F1 values due to reduced minority-class performance. These results confirm that macro-F1 provides a more reliable measure of model effectiveness in imbalanced multi-class conversational datasets than accuracy alone.

Table 2. Model-wise Performance Comparison

Model	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)
BERT	91.2	88.5	90.7
GPT	89.4	84.1	88.3
XLNet	90.1	86.9	89.6
Mistral	88.7	83.8	87.9
LLaMA	90.5	87.6	89.9

This table provides a consolidated numerical comparison and supports the visual trends observed in Figure 6.

8. DISCUSSION

This section interprets the experimental results and explains the observed performance differences among the evaluated transformer models. The discussion focuses on model architecture characteristics, class imbalance behavior, and practical implications for context-aware conversational intelligence systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

The results demonstrate that BERT consistently outperforms other models in terms of macro-F1 **score**, indicating superior handling of minority classes. This performance can be attributed to BERT's bidirectional encoder architecture, which enables richer contextual representations for discriminative classification tasks. By jointly modeling left and right context, BERT effectively captures subtle semantic differences between classes, reducing misclassification in overlapping categories.In contrast, GPT exhibits weaker macro-F1 performance, despite achieving competitive accuracy. This behavior highlights the limitations of autoregressive models in multi-class classification settings. GPT's training objective prioritizes next-token prediction rather than explicit class separation, which leads to reduced recall for underrepresented classes. These findings confirm that accuracy-centric evaluation is insufficient for assessing conversational systems operating under imbalanced data conditions.LLaMA demonstrates a strong balance between performance and efficiency, achieving macro-F1 scores close to BERT while maintaining lower computational complexity. This result suggests that parameter-efficient large language models can serve as viable alternatives for real-world conversational systems, particularly when resource constraints are present. XLNet achieves moderate performance improvements over GPT, benefiting from its permutation-based training strategy, while Mistral shows competitive efficiency but lower minority-class recall.Overall, the discussion reinforces the importance of selecting evaluation metrics aligned with deployment requirements. For conversational intelligence systems where rare intents are critical, models optimized for macro-F1 performance are more suitable than those optimized solely for accuracy.

9. CHALLENGES AND LIMITATIONS

Despite the comprehensive evaluation framework, several challenges and limitations remain. First, transformer-based models require substantial computational resources for fine-tuning and inference, which may limit their deployment in resource-constrained environments. Second, the evaluation is conducted on benchmark datasets that, while widely used, may not fully capture the complexity of real-world conversational interactions.

Additionally, the study focuses primarily on quantitative evaluation metrics derived from confusion matrices. Human-centered evaluation, such as user satisfaction and response appropriateness, is not considered and represents an important area for future research. Finally, the results may vary across domains, and domain-specific fine-tuning could further influence model behavior.

10. ETHICAL CONSIDERATIONS

Ethical considerations are critical in the development and deployment of conversational AI systems. Transformer-based models may inherit biases present in training data, leading to unfair or discriminatory behavior. Ensuring fairness across classes and user groups is essential, particularly in applications involving sensitive information or decision-making.Data privacy is another key concern, as conversational systems often process personal or confidential user data. Proper data anonymization, secure storage, and compliance with data protection regulations are necessary to ensure responsible deployment. Transparency in model behavior and evaluation practices also contributes to ethical AI development by enabling accountability and informed decision-making.

11. CONCLUSION & FUTURE WORK

This paper presents a comprehensive comparative evaluation of transformer-based models for context-aware conversational intelligence in imbalanced multi-class settings. By analysing GPT, BERT, XLNet, Mistral and LLaMA within a unified experimental framework, the study demonstrates that accuracy alone isn't enough for evaluating conversational systems on real-world datasets with uneven class distributions.A confusion matrix-driven evaluation strategy was used to calculate precision, recall and F1-based metrics, particularly focusing on the macro-F1 score. The results showed BERT consistently outperformed the others in macro-F1 performance due to its bidirectional contextual representation, which effectively handles minority classes. LLaMA offered a strong balance between performance and computational efficiency, making it a practical choice for real-world applications. In contrast, autoregressive models like GPT achieved competitive accuracy but reduced recall for minority classes, highlighting the limitations of accuracy-centric evaluation.The findings confirm that model selection for conversational intelligence systems should prioritise class-wise performance metrics that align with application requirements, especially when rare intents are crucial. The proposed evaluation framework provides practical guidance for selecting transformer models based on robust and interpretable performance indicators rather than just aggregate accuracy.Future work could extend this study by incorporating domain-specific fine-tuning, long-context transformer architectures and explainable AI techniques to improve interpretability. Integrating human-centred evaluation and real-world conversational datasets would further strengthen the applicability of the proposed framework.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

12. REFERENCES

[1] S. S. Wang, Y. X. Liu, and Z. G. Wang. (2021). "Confusion Matrix in Deep Learning: A Review of Metrics and Applications." *International Journal of Computational Intelligence*, 9(3), 231-244.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. (2017). "Attention is All You Need." *Advances in Neural Information Processing Systems*, 30, 5998-6008.
[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. (2020). "Language Models are Few-Shot Learners." *arXiv preprint arXiv:2005.14165*.
[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.
[5] G. H. Shihab and M. A. Farag. (2020). "Evaluation of Chatbot Implementation in Customer Service Using Confusion Matrix Analysis." *Journal of Business Informatics*, 5(1), 44-51.
[6] M. A. Young, J. A. Thomason, and H. L. Jin. (2021). "Conversational".
Chatbots with Transformer-Based Architectures: A Survey." *IEEE Transactions on Computational Social Systems*, 8(2), 247-260.
[7] P. R. Gupta, S. Dhankhar, and R. Kaushik. (2019). "Sentiment Analysis and Chatbot Conversation: A Literature Review." *Journal of Information Science and Engineering*, 35(5), 1030-1045.
[8] D. Zhang and J. Zhao. (2019). "Confusion Matrix and Its Application in Deep Learning for Image Classification." *IEEE Transactions on Image Processing*, 28(12), 5645-5652.
[9] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon. (2019). "Unified Language Model Pre-training for Natural Language Understanding and Generation." *arXiv preprint arXiv:1905.03197*.
[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. (2019). "Language Models are Unsupervised Multitask Learners." *OpenAI Blog*.
[11] K. Cho, B. Van Merriënboer, Ç. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. (2014). "Learning Phrase Representations using Recurrent Neural Networks."
An RNN Encoder-Decoder for Statistical Machine Translation, published as an *arXiv preprint: arXiv:1406.1078*.
[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). "RoBERTa: A Robustly Optimised BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692*.
[13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le (2019). "XLNet: Generalised Autoregressive Pretraining for Language Understanding." Advances in Neural Information Processing Systems, 32, 5754-5764.
[14] K. Clark, M. Luong, Q. V. Le, and C. D. Manning (2020). "Electra: Pre-training Text Encoders as Discriminators Rather Than Generators." *arXiv preprint arXiv:2003.10555*.
[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). "Natural Language Processing (Almost) from Scratch." *Journal of Machine Learning Research*, 12, 2493-2537.
[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and P.
Cistac et al. (2020) published "Transformers: State-of-the-Art Natural Language Processing*" in the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38-45.
[17] X. Glorot, A. Bordes, and Y. Bengio (2011) presented "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach*" at the 28th International Conference on Machine Learning (ICML-11). Their work is summarised in the proceedings*, pages 513-520.
[18] D. P. Kingma and J. Ba (2015) introduced "Adam: A Method for Stochastic Optimisation*" at the 3rd International Conference on Learning Representations (ICLR 2015). This method revolutionised the field of stochastic optimisation*.
[19] F. Chollet (2017) showcased "Xception: Deep Learning with Depthwise Separable Convolutions*" at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. Xception significantly improved the performance of deep learning models.
[20] A. Vaswani, L. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017) presented "Attention is All You Need*" at Advances in Neural Information Processing Systems (NIPS). This groundbreaking work introduced the attention mechanism*, revolutionising natural language processing.
[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, et al. (2016) introduced "TensorFlow: A System for Large-Scale Machine Learning" at the International Conference on Machine Learning (ICML). TensorFlow has become a widely used platform for large-scale machine learning projects. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 2016)*, pages 265-283.
[22] X. Zhang, J. Zhao, and Y. LeCun. (2015). "Character-Level Convolutional Networks for Text Classification." *In Advances in Neural Information Processing Systems (NIPS)*, pages 649-657.
[23] M. D. Zeiler and R. Fergus. (2014). "Visualising and Understanding Convolutional Networks." *In European Conference on Computer Vision (ECCV 2014)*, pages 818-833.
[24] B. Han, S. Goel, and S. H. Kang. (2020). "Deep Learning for Multi-Class Classification: A Case Study in Automated Biomedical Image Analysis." *Frontiers in Artificial Intelligence*, volume 3, pages 10-23.
[25] C. Szegedy, V. Vanhoucke, Z. Zhong, and A. Rabinovich. (2016). "Rethinking the Inception Architecture for Computer Vision." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pages 2818-2826.*
[26] T. Mikolov, K. Chen, G. Corrado and J. Dean. (2013). "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*.
[27] M. J. T. Reinders and W. J. D. M. A. De Haas. (2022). "Multi-Class Classification for Deep Learning Models." *Springer International Publishing*.
[28] Y. Bengio. (2009). "Learning Deep Architectures for AI." *Foundations and Trends in Machine Learning*, 2(1), 1-127.
[29] J. Schulman, P. Abbeel, D. P. Kingma, and J. Ho. (2015). "Trust Region Policy Optimisation." *In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1889-1897.