ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2025; 4; 1-19

A Transformer-Based Cognitive Classification Framework for Integrating Bloom's Taxonomy into Adaptive Testing Systems

[a]Isaac Olawale Ifinju* , **[b]Mayowa O. Ogunjimi** & **[c]Timothy Obasuyi Oghogho**

[ab] Department of Social Sciences Education, Faculty of Education, University of Ilorin, Ilorin, Nigeria
[c]Department of Information System, East Tennessee State University, Tennessee USA

Abstract

This study proposes a transformer-based machine learning framework for automated classification of test items according to Bloom's taxonomy of cognitive processes, aiming to enhance cognitive alignment in adaptive testing systems. Manual classification of assessment items is often time-intensive and susceptible to subjectivity, posing challenges for large-scale and real-time adaptive testing environments. To address this limitation, the study employs Natural Language Processing techniques, specifically fine-tuning a BERT-based model, to classify Economics test items into six cognitive levels. A balanced dataset comprising 3,000 expert-annotated items (Cohen's $\varkappa = 0.87$) was used, with stratified training, validation, and testing procedures to ensure robustness. Model performance was evaluated using accuracy, precision, recall, and F1-score, alongside stratified 5-fold cross-validation and comparative analysis with Logistic Regression and Support Vector Machine baselines. The proposed model achieved an accuracy of 97.7% on the held-out test set and a cross-validated mean accuracy of 96.7% (SD = 0.5), substantially outperforming baseline models. While classification performance was consistently high across most categories, relatively lower recall for the "Understand" level suggests inherent semantic overlap between adjacent cognitive processes. The findings demonstrate the potential of transformer-based models to capture nuanced cognitive demand in assessment items. Importantly, the study conceptualises this classification framework as a complementary layer to Item Response Theory-based adaptive testing systems, enabling item selection that accounts for both psychometric properties and cognitive complexity. The study contributes a scalable approach to improving the consistency and pedagogical alignment of item classification, while highlighting the need for further validation using diverse and imbalanced datasets and for integration into operational adaptive testing environments.

Keywords: machine learning, bloom's taxonomy, adaptive testing, nlp, educational data mining, transformers

* Corresponding author.

Department of Social Sciences Education, Faculty of Education, University of Ilorin, Nigeria, e-mail: olawaleifinju@gmail.com

To cite this article:

Ifinju, I. O., Ogunjimi, M. O. & Oghogho, T. O. (2026). A Transformer-Based Cognitive Classification Framework for Integrating Bloom's Taxonomy into Adaptive Testing Systems. *Journal of Computerised Adaptive Testing in Africa, 4*, 1-19.

Received: 3 September 2025
Accepted: 70 December 2025
Published: 23 December 2025

INTRODUCTION

Effectively assessing students' learning outcomes remains a central concern in educational research and practice, particularly in contexts where assessment must capture both the breadth and depth of learners' cognitive development. Among the frameworks guiding the design and evaluation of assessment tasks, Bloom's Taxonomy remains one of the most widely adopted due to its structured classification of cognitive processes, ranging from lower-order skills such as remembering to higher-order processes including analysing, evaluating, and creating (Banda et al., 2023). By providing a hierarchical representation of cognitive demand, Bloom's taxonomy supports the development of assessments that extend beyond factual recall to encompass complex reasoning and problem-solving. Consequently, accurate classification of test items by cognitive level is essential to ensuring that assessments validly represent intended learning outcomes.

Despite its widespread use, the practical implementation of Bloom's taxonomy in assessment design remains largely dependent on manual classification by subject-matter experts. As noted by Meissner et al. (2020), this process is inherently time-intensive and subject to inconsistency, as different experts may assign divergent cognitive levels to the same item. Such variability introduces reliability concerns and threatens the consistency of item banks, particularly in large-scale assessment systems. These challenges are further amplified in adaptive testing environments, where the real-time selection of items requires rapid, consistent, and scalable classification mechanisms. Without reliable cognitive tagging, adaptive systems risk selecting items that are psychometrically appropriate but cognitively misaligned with intended learning objectives.

Recent advances in Machine Learning (ML) and Natural Language Processing (NLP) present promising avenues for addressing these limitations by enabling the automated classification of test items based on their linguistic and semantic properties. Early studies have explored the use of traditional machine learning algorithms and basic NLP techniques for Bloom's taxonomy classification (Asthana et al., 2024; Gavhane & Pagare, 2025). While these approaches demonstrate initial feasibility, they are largely constrained by their reliance on handcrafted features and shallow textual representations, limiting their capacity to capture the nuanced semantic and contextual cues that distinguish cognitive levels. More recently, transformer-based models have achieved substantial improvements in text classification tasks by learning contextualised language representations. However, their application to Bloom's taxonomy classification remains underdeveloped, particularly

ISSN: 2790-4407 www. jocatia.acata.org

within structured assessment contexts such as adaptive testing systems and in underrepresented educational settings, including many African contexts where domain-specific datasets and scalable computational frameworks are still emerging.

At the same time, modern adaptive testing systems are predominantly grounded in psychometric frameworks such as Item Response Theory (IRT), which optimise item selection based on statistical properties including difficulty, discrimination, and information (An & Yung, 2014). While these models are highly effective in estimating learner ability and maximising measurement precision, they do not explicitly account for the cognitive processes required to engage with test items. As a result, adaptive testing systems may achieve psychometric efficiency while failing to ensure balanced representation across different levels of cognitive demand. This limitation has important implications for construct validity, as assessments that neglect higher-order cognitive processes may provide an incomplete representation of learners' capabilities.

Addressing this limitation requires a complementary approach that integrates cognitive classification into psychometric frameworks, thereby enabling adaptive testing systems to consider both measurement precision and cognitive alignment. In this regard, automated classification of test items based on Bloom's taxonomy can serve as a critical layer within adaptive systems, supporting the organisation of item banks and informing constraint-based item selection strategies that ensure appropriate cognitive coverage.

Against this backdrop, the present study proposes a transformer-based machine learning framework for the automated classification of test items according to Bloom's taxonomy, using a domain-specific dataset of Economics items from Nigeria. By leveraging the contextual modelling capabilities of transformer architectures, the study seeks to enhance the accuracy, consistency, and scalability of cognitive-level classification. More importantly, the study conceptualises this classification framework as a complementary component to IRT-based adaptive testing systems, thereby contributing to the integration of cognitive and psychometric perspectives in assessment design.

Specifically, the study pursues the following objectives:

i. To develop a transformer-based machine learning framework for classifying test items according to Bloom's taxonomy.

ii. To evaluate the performance of the proposed model using standard classification metrics and cross-validation procedures.

iii. To examine the alignment between actual and predicted cognitive classifications through graphical and distributional analysis.

iv. To compare the performance of the transformer-based model with selected traditional machine learning approaches.

LITERATURE REVIEW

Bloom's Taxonomy has long served as a foundational framework for structuring educational objectives and assessment tasks. Since its introduction and subsequent revision, it has been widely applied in curriculum design, instructional planning, and assessment development to ensure coverage of diverse cognitive processes (Forehand, 2010; Chandio et al., 2016). By organising learning outcomes into hierarchical cognitive levels, the taxonomy provides a conceptual basis for designing assessments that capture both lower- and higher-order thinking skills. However, despite its theoretical robustness and widespread adoption, the operationalisation of Bloom's taxonomy in test item classification remains predominantly manual. This reliance on expert judgement raises concerns about scalability, consistency, and inter-rater variability, particularly in large-scale, adaptive testing contexts where rapid and reliable item categorisation is essential.

In response to these limitations, research in Educational Data Mining (EDM) and learning analytics has increasingly explored automated approaches to assessment processes (Du et al., 2020). Early efforts in automated item classification were grounded in rule-based systems and traditional machine learning techniques, including logistic regression, decision trees, and support vector machines (AlKhuzaey et al., 2021). While these approaches offered initial improvements in efficiency, their reliance on handcrafted features constrained their ability to capture the deeper semantic and contextual characteristics embedded in assessment items. As Bloom's taxonomy classification requires sensitivity to subtle variations in cognitive demand, models based on surface-level representations often struggle to distinguish between adjacent cognitive levels, particularly within lower-order categories where linguistic cues are less explicit.

ISSN: 2790-4407 www. jocatia.acata.org

The introduction of word embedding techniques, such as Word2Vec and GloVe, marked a significant advancement by enabling models to represent semantic relationships between words as continuous vectors. These approaches improved classification performance by capturing lexical similarity and contextual proximity. However, their context-independent nature limits their ability to resolve ambiguity in meaning, particularly in educational texts where identical terms may signal different cognitive processes depending on context. Consequently, while word embeddings represent an important step forward, they remain insufficient for accurately modelling the nuanced cognitive distinctions required for Bloom's taxonomy classification.

More recently, transformer-based models, including BERT, RoBERTa, and GPT, have transformed the landscape of Natural Language Processing by enabling the learning of contextualised language representations (Dogra et al., 2022; Cunha et al., 2023). Unlike earlier approaches, these models capture both syntactic and semantic dependencies within text, allowing for more precise interpretation of meaning. In educational applications, transformer models have demonstrated strong performance in tasks such as automated grading, feedback generation, and recommendation systems (Sung et al., 2019; Lu & Cutumisu, 2021; Madhavi et al., 2023). Emerging evidence further suggests that these models are particularly effective in identifying subtle differences in cognitive demand, making them well-suited for Bloom's taxonomy classification (Bhopale & Tiwari, 2024). Despite these advances, existing studies have largely focused on improving classification accuracy in isolation, with limited attention to integrating these models into structured assessment systems. In particular, there is a lack of research examining how automated cognitive classification can be operationalised within adaptive testing environments, where item selection must balance multiple criteria. This gap is especially pronounced in developing educational contexts, including many African systems, where domain-specific datasets and scalable computational frameworks for assessment remain underdeveloped. To synthesise the evolution of these approaches and clarify their relative strengths and limitations, Table 1a presents a comparative analysis of NLP techniques used for Bloom's taxonomy classification.

Table 1a: Comparative Analysis of NLP Approaches for Bloom's Taxonomy Classification

Approach	**Key Studies**	**Strengths**	**Limitations**	**Implication for Current Study**
Rule-Based Systems	Early EDM studies	Simple to implement; interpretable	Rigid; cannot generalise; poor scalability	Not suitable for large-scale adaptive systems
Traditional ML (LR, SVM, DT)	AlKhuzaey et al. (2021)	Efficient; works with small datasets	Relies on handcrafted features; shallow semantic understanding	Limited ability to distinguish higher-order cognitive levels
Word Embeddings (Word2Vec, GloVe)	Du et al. (2020)	Captures semantic similarity between words	Context-independent; struggles with ambiguity	Insufficient for nuanced classification tasks
Transformer Models (BERT, RoBERTa, GPT)	Dogra et al. (2022); Cunha et al. (2023); Bhopale & Tiwari (2024)	Context-aware; captures deep semantic and syntactic relationships; high accuracy	Computationally intensive; requires large datasets; limited interpretability	Most suitable for modelling cognitive complexity in assessment items

The progression illustrated in Table 1a reflects a clear shift from rule-based and shallow machine learning approaches toward deep contextual models capable of capturing complex semantic relationships. While transformer-based models demonstrate superior performance in classification tasks, the literature reveals a critical gap: existing studies largely prioritise predictive accuracy without adequately addressing how such models can be operationalised within structured assessment systems. In particular, there is limited research on how automated cognitive classification can be integrated into adaptive testing environments that must simultaneously satisfy psychometric and pedagogical constraints.

In parallel with advances in NLP, Item Response Theory (IRT) remains central to modern assessment, particularly in the development of Computerised Adaptive Testing (CAT) systems. IRT

ISSN: 2790-4407 www. jocatia.acata.org

models, including the Rasch model, estimate learner ability based on item parameters such as difficulty and discrimination, enabling dynamic item selection that maximises measurement precision (An & Yung, 2014). However, a critical limitation of IRT-based systems is their exclusive focus on psychometric properties, without explicit consideration of the cognitive processes required by test items. As a result, adaptive tests may optimise statistical efficiency while failing to ensure balanced representation across different levels of cognitive demand, raising concerns about construct validity.

To address this limitation, recent scholarship has begun to emphasise the need to integrate cognitive and psychometric perspectives into assessment design. Within this emerging perspective, automated classification of test items based on Bloom's taxonomy can serve as a complementary layer to IRT-based systems. Specifically, transformer-based models can be used to assign cognitive-level labels to assessment items, while IRT models provide information on item difficulty and discrimination. The integration of these components enables a dual-informed assessment framework in which item selection is guided not only by statistical efficiency but also by cognitive alignment.

This integration can be operationalised in two key ways. First, at the item bank development level, classified items can be organised by cognitive level, ensuring that the assessment pool reflects a balanced distribution of cognitive demand. Second, within adaptive testing algorithms, constraint-based item selection strategies can be implemented to ensure that selected items maximise information while maintaining appropriate cognitive coverage. Such an approach supports the development of adaptive systems that are both psychometrically robust and pedagogically meaningful. By positioning NLP-based Bloom's taxonomy classification as a supporting mechanism rather than a replacement for IRT, the present study addresses a critical gap in literature. It moves beyond isolated applications of machine learning and psychometric modelling, proposing an integrated framework that enhances both measurement precision and cognitive validity in adaptive testing systems.

METHODOLOGY

This study adopts a machine learning approach to automatically classify test items into Bloom's Taxonomy cognitive levels, with the broader aim of supporting cognitively informed adaptive testing systems. The methodological framework is structured into five sequential stages: data preparation, preprocessing, model development, evaluation, and prediction/deployment. This pipeline ensures

both analytical rigour and reproducibility while aligning with best practices in Natural Language Processing and educational data mining.

Data Preparation

A balanced dataset comprising 3,000 test items (500 per Bloom's taxonomy level) was constructed to ensure equal representation across all six cognitive categories (Remember, Understand, Apply, Analyse, Evaluate, Create). This design mitigates model bias toward dominant classes and facilitates stable learning across categories, although it is acknowledged that real-world assessment datasets often exhibit class imbalance. Consequently, the balanced design is intended to support controlled model development rather than to replicate operational testing conditions. To assess the dataset's reliability, two subject-matter experts independently annotated all test items using Bloom's taxonomy. Inter-rater reliability was assessed using Cohen's Kappa coefficient, yielding $\kappa = 0.87$, which indicates substantial agreement. Discrepancies between raters were resolved through a structured consensus process to produce the final ground truth labels used for model training. The dataset was partitioned into training (n = 2,400), validation (n = 300), and test (n = 300) subsets using stratified sampling, ensuring proportional representation of all cognitive levels across splits. This approach supports model generalisation while preventing distributional bias.

Preprocessing

Before model training, the textual content of test items underwent systematic preprocessing to enhance data quality and consistency. This included removing non-informative symbols, standardising text formats, and normalising punctuation. Tokenisation was performed using a transformer-compatible tokenizer, which converts input text into subword units suitable for deep contextual models. To prevent data leakage, strict separation was maintained across training, validation, and test datasets, ensuring that no overlapping or duplicated items appeared across splits. Stratified sampling procedures were further applied to preserve class balance without introducing redundancy. These steps are critical for ensuring the validity and integrity of model evaluation.

Model Development

The classification model was implemented using the Hugging Face Transformers library, with a BERT-base-uncased architecture fine-tuned for multi-class text classification. This architecture was

ISSN: 2790-4407 www. jocatia.acata.org

selected due to its ability to capture contextual semantic relationships, which are essential for distinguishing subtle variations in cognitive demand across Bloom's taxonomy levels. Input sequences were tokenised using the BERT tokeniser, with a maximum sequence length of 128 tokens and truncation and padding as required. The model was fine-tuned with a learning rate of 2×10^{-5} and a batch size of 16, and trained for three epochs. Optimisation was performed using the AdamW optimiser, with cross-entropy loss applied for multi-class classification. To ensure experimental reproducibility, a fixed random seed (42) was used across all training procedures. Model training was conducted on a GPU-enabled system to improve computational efficiency. Stratified sampling was maintained throughout the training process to preserve class balance and ensure stable parameter estimation.

Evaluation

Model performance was evaluated using a held-out test set of 300 items, employing standard classification metrics including accuracy, precision, recall, and F1-score for each cognitive category. Both macro-averaged and weighted metrics were computed to provide a comprehensive assessment of model performance across balanced classes. To assess robustness and mitigate overfitting, a stratified 5-fold cross-validation procedure was conducted. The dataset was partitioned into five folds while preserving class distribution, and the model was iteratively trained and evaluated on each fold. Mean performance metrics and corresponding standard deviations were computed to evaluate model stability and generalisability. For comparative analysis, baseline classifiers including Logistic Regression and Support Vector Machine (SVM) were implemented. All models were trained and evaluated under identical conditions, including data splits and preprocessing pipelines, to ensure fairness in comparison. To statistically assess performance differences between the transformer-based model and baseline models, McNemar's test was applied to paired classification outputs, enabling evaluation of whether observed improvements were statistically significant.

Prediction and Deployment

In the final stage, the trained model was applied to classify unseen test items into Bloom's taxonomy levels, demonstrating its applicability in real-world assessment scenarios where new items are continuously generated. The automated classification output provides cognitive-level metadata that can be integrated into assessment workflows, particularly within adaptive testing systems. While Item Response Theory (IRT) remains central to adaptive testing through its estimation of item

parameters and learner ability, the present study focuses on developing a complementary cognitive classification layer. By assigning cognitive labels to test items, the framework supports the organisation of item banks and enables the incorporation of cognitive constraints into item selection algorithms. This facilitates the development of adaptive testing systems that are not only psychometrically efficient but also aligned with intended cognitive learning outcomes.

RESULT

Objective 1: Development of the Transformer-Based Classification Framework

The proposed transformer-based machine learning framework for classifying test items according to Bloom's taxonomy is presented in Figure 1. The framework outlines a structured pipeline comprising data preparation, preprocessing, model fine-tuning, evaluation, and deployment. This sequential architecture ensures a reproducible and scalable approach to automated cognitive classification.

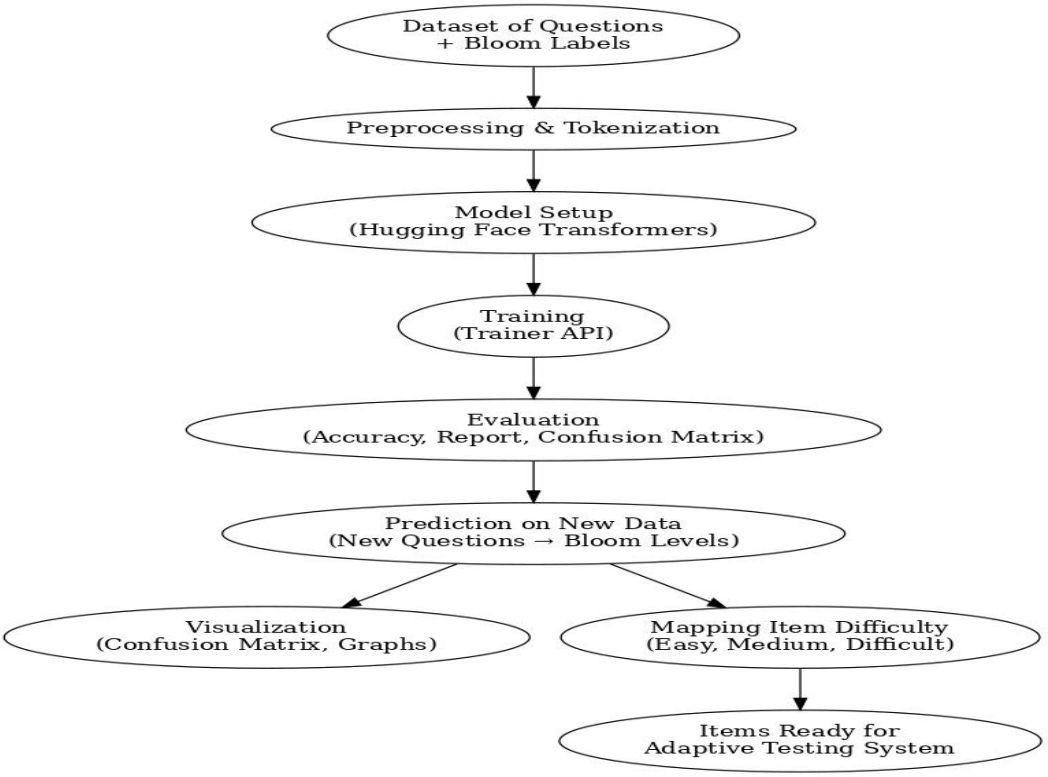

Figure 1: Transformer-Based Framework Workflow (Source: Author's Design, 2025)

ISSN: 2790-4407

As illustrated in Figure 1, the classification model was implemented using the BERT-base-uncased architecture from the Hugging Face Transformers library. Input text was tokenised using the corresponding BERT tokenizer, with a maximum sequence length of 128 tokens, applying truncation and padding as required. The model was fine-tuned using a learning rate of 2×10^{-5}, a batch size of 16, and trained over three epochs. Optimisation was performed using the AdamW optimiser with a cross-entropy loss function. A stratified data splitting strategy was employed to maintain class distribution across training, validation, and test sets. To ensure reproducibility, a fixed random seed (42) was applied throughout the training process. Model training was conducted in a GPU-enabled environment to enhance computational efficiency. This framework provides the computational foundation for integrating automated cognitive classification into adaptive testing systems.

Objective 2: Performance of the Transformer-Based Model

The classification performance of the transformer-based model on the held-out test set is presented in Figure 2.

The result below presents the classification performance of the transformer-based model on a held-out test set

```
Overall Bloom-Level Metrics
----------------------------
Accuracy : 0.977
Precision: 0.978
Recall   : 0.977
F1 Score : 0.976

Per-Class Bloom-Level Report
----------------------------
              precision    recall  f1-score   support

     Analyze       0.98      1.00      0.99        50
       Apply       1.00      1.00      1.00        50
      Create       1.00      1.00      1.00        50
    Evaluate       0.98      1.00      0.99        50
    Remember       0.91      1.00      0.95        50
  Understand       1.00      0.86      0.92        50

    accuracy                           0.98       300
   macro avg       0.98      0.98      0.98       300
weighted avg       0.98      0.98      0.98       300
```

Figure 2: Classification Performance of the Framework

The model achieved an overall accuracy of 97.7%, with both macro-averaged and weighted-average F1-scores of approximately 0.98, indicating consistently high performance across all Bloom's taxonomy levels. At the class level, the model demonstrated near-perfect performance for Apply and Create (F1 = 1.00), and strong performance for Analyze and Evaluate (F1 = 0.99).

Relatively lower performance was observed for Remember (F1 = 0.95) and Understand (F1 = 0.92), with the latter exhibiting the lowest recall (0.86). This pattern indicates that misclassification occurs primarily between adjacent lower-order cognitive levels, suggesting challenges in distinguishing semantically similar categories.

To assess model robustness, a stratified 5-fold cross-validation procedure was conducted. The model achieved a mean accuracy of 96.7% (SD = 0.5) across folds, indicating stable performance and consistent generalisation across different data partitions.

Objective 3: Comparison of Actual and Predicted Classifications

A graphical comparison of actual and predicted Bloom's taxonomy classifications is presented in Figure 3.

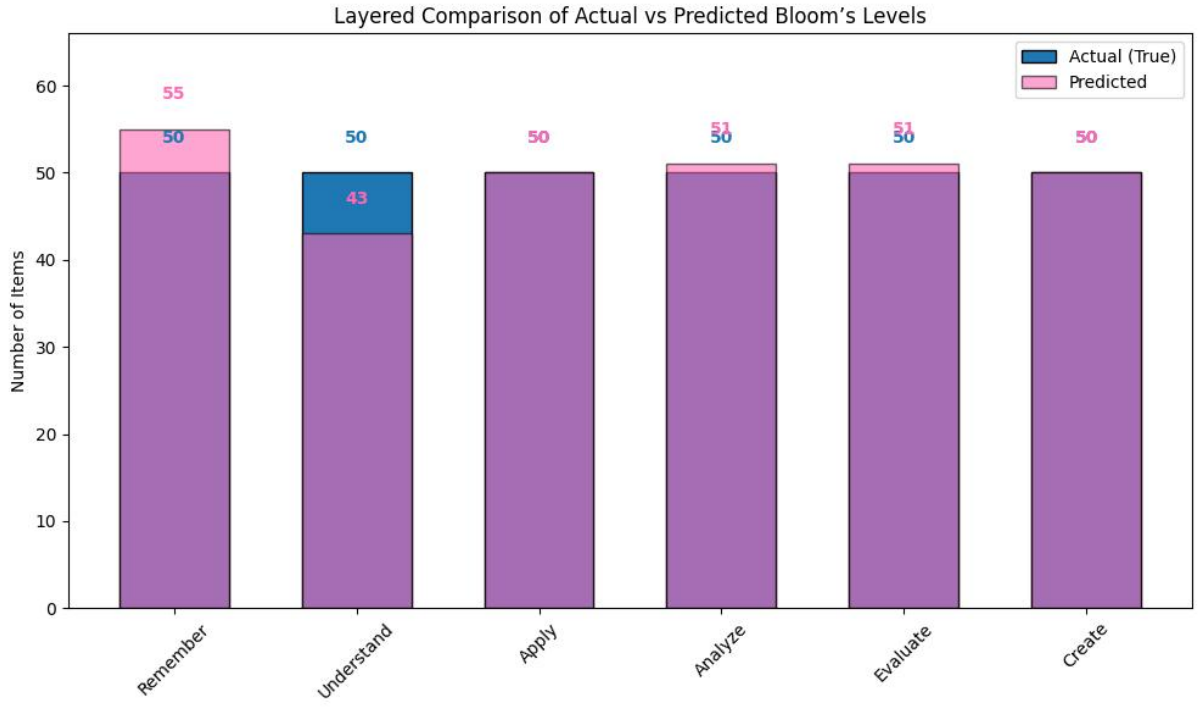

Figure 3: Layered Comparison of Actual and Predicted Bloom's Taxonomy Levels

ISSN: 2790-4407

Figure 3 shows a strong correspondence between actual and predicted distributions across all six cognitive categories. The predicted counts closely align with the true distribution, confirming the model's high classification accuracy.

Minor discrepancies are observed in the Remember (predicted = 55; actual = 50) and Understand (predicted = 43; actual = 50) categories. These deviations indicate limited misclassification between closely related cognitive levels. In contrast, predictions for higher-order categories (Apply, Analyze, Evaluate, and Create) closely match actual values, reflecting the model's strong capability to distinguish cognitively complex items.

Objective 4: Comparison with Traditional Machine Learning Models

Table 1b presents a comparative evaluation of the transformer-based model against baseline machine learning classifiers.

Table 1b. Model Performance Summary

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	61.4	32.4	35.1	50.2
SVM	76.4	74.3	76.2	76.3
Transformer (Proposed)	97.7	97.8	97.7	97.6

As shown in Table 1b, the transformer-based model substantially outperformed traditional machine learning approaches across all evaluation metrics. Logistic Regression exhibited limited performance, particularly in precision and recall, indicating its inability to capture the semantic complexity required for cognitive classification. The SVM model demonstrated moderate performance; however, it remained considerably lower than that of the transformer-based model.

To determine whether these differences were statistically significant, McNemar's test was conducted on paired classification outputs. The results indicated that the performance improvement achieved by the transformer-based model over the baseline models was statistically significant ($p < 0.05$), confirming that the observed gains are not attributable to random variation.

DISCUSSION

The findings of this study provide important insights into the application of transformer-based models for cognitive classification and their potential to advance adaptive testing systems. First, the development of the proposed framework demonstrates the feasibility of constructing a structured, reproducible pipeline for the automated classification of test items according to Bloom's taxonomy. The integration of data preparation, preprocessing, model fine-tuning, evaluation, and deployment reflects current best practices in Natural Language Processing, particularly the use of transformer architectures to model contextual meaning. Consistent with prior research (Zaidi et al., 2018), adopting a BERT-based model enabled the capture of nuanced semantic relationships embedded in assessment items, which are critical for distinguishing between cognitive levels. Beyond its technical implementation, the framework contributes conceptually by positioning cognitive classification as an operational layer that can be integrated into adaptive testing systems. In this regard, the study extends existing work by moving from standalone classification towards a system-oriented perspective in which cognitive alignment complements psychometric modelling rather than replacing it.

The second set of findings, relating to model performance, indicates that the transformer-based approach achieves consistently high classification accuracy across Bloom's taxonomy levels. The observed overall accuracy (97.7%) and strong macro- and weighted-average F1-scores suggest that the model effectively captures the linguistic features associated with different cognitive processes. However, a more nuanced interpretation of class-level performance reveals important patterns. The model performs most effectively on higher-order cognitive categories such as Apply, Analyse, Evaluate, and Create, while lower recall is observed at lower-order levels, particularly Understand. This pattern is theoretically meaningful and aligns with the hierarchical structure of Bloom's taxonomy, where lower-order categories often share overlapping semantic characteristics (Asthana et al., 2024). Unlike higher-order tasks, which tend to involve more explicit linguistic markers of reasoning or evaluation, lower-order tasks frequently rely on similar lexical constructions, making them inherently more difficult to distinguish. The consistency of model performance across cross-validation folds further supports the robustness of these findings, indicating that the model generalises well within the dataset. Nevertheless, the controlled nature of the balanced dataset

ISSN: 2790-4407 www. jocatia.acata.org

suggests that future research should examine performance under more realistic conditions, including imbalanced and cross-domain datasets.

The graphical comparison between actual and predicted classifications provides additional evidence of the model's effectiveness in capturing the cognitive structure of assessment items. The close alignment between observed and predicted distributions across all categories indicates that the model is not only accurate at the individual item level but also preserves the overall cognitive profile of the dataset. Minor discrepancies observed in the Remember and Understand categories reinforce the earlier finding that semantic proximity between adjacent cognitive levels presents classification challenges. Importantly, such misclassifications are not unique to automated systems; they reflect inherent ambiguities in item construction and have been reported in prior studies involving human raters and machine learning models alike (Besar et al., 2025). From a measurement perspective, this suggests that classification uncertainty at lower cognitive levels may be partly attributable to construct overlap rather than purely algorithmic limitations.

The comparative analysis further demonstrates the substantial advantage of transformer-based models over traditional machine learning approaches. While Logistic Regression and Support Vector Machine models showed limited and moderate performance respectively, the transformer model achieved markedly higher scores across all evaluation metrics. This performance gap can be attributed to transformer architectures' ability to model contextual dependencies and semantic relationships in text, which are essential for interpreting cognitive intent in assessment items (Banda et al., 2023). Traditional models, by contrast, rely on surface-level representations that are insufficient for capturing the complexity of language associated with different cognitive processes. The statistical significance of these differences, as confirmed by McNemar's test, reinforces the robustness of the observed improvements and supports the argument that transformer-based approaches represent a meaningful advancement in automated cognitive classification.

Beyond performance improvements, the findings have broader implications for the design and implementation of adaptive testing systems. By enabling automated classification of items by cognitive demand, the proposed framework opens the possibility of integrating cognitive constraints into item selection processes. This has important implications for construct validity, as it allows adaptive tests to move beyond purely psychometric optimisation towards a more balanced representation of cognitive processes. In this sense, the study contributes to an emerging perspective in educational measurement that emphasises the integration of cognitive and psychometric

dimensions in assessment design. The framework developed in this study provides an initial step toward such integration by demonstrating how transformer-based models can generate cognitive-level metadata that complements IRT-based item parameters.

However, several limitations should be acknowledged. Using a balanced dataset, while beneficial for controlled model training, may not fully reflect the distribution of cognitive levels in real-world assessment contexts. Additionally, the study focuses on a single subject domain, which may limit the generalisability of findings across disciplines. The model's performance in distinguishing lower-order cognitive levels also highlights the need for further refinement, potentially through the incorporation of additional contextual features or hybrid modelling approaches. Future research should therefore explore cross-domain validation, the impact of dataset imbalance, and the integration of cognitive classification into operational adaptive testing systems through simulation or live deployment.

CONCLUSION AND RECOMMENDATIONS

This study developed and evaluated a transformer-based machine learning framework for automated classification of test items according to Bloom's taxonomy, supporting cognitively informed adaptive testing systems. The findings demonstrate that the proposed framework achieves high and consistent classification performance across cognitive levels, as evidenced by strong evaluation metrics and stable cross-validation results. In particular, the model demonstrated robust performance in identifying higher-order cognitive processes, underscoring the effectiveness of transformer-based architectures in capturing the complex semantic and contextual features embedded in assessment items.

At the same time, the observed misclassifications between lower-order cognitive levels, particularly Remember and Understand, reflect inherent conceptual overlaps within Bloom's taxonomy rather than purely algorithmic limitations. This finding reinforces the need to interpret classification outcomes within the broader theoretical context of cognitive frameworks, where category boundaries are not always sharply defined.

Beyond its empirical performance, the study contributes to the literature by advancing a conceptual integration of cognitive classification and psychometric modelling. By positioning automated Bloom's taxonomy classification as a complementary layer to Item Response Theory-based systems, the framework provides a pathway toward adaptive testing systems that are not only statistically efficient but also cognitively aligned with intended learning outcomes. This dual-informed

ISSN: 2790-4407

perspective represents an important step toward enhancing construct validity in technology-driven assessment.

In light of these contributions, several recommendations are proposed. First, future research should focus on integrating the proposed framework into large-scale adaptive testing systems to enable real-time classification and cognitively informed item selection. Such integration would enable empirical evaluation of the framework's impact on measurement precision, test efficiency, and cognitive coverage in operational environments. Second, further studies should examine model performance under imbalanced data conditions to better reflect real-world assessment scenarios, where certain cognitive levels may be underrepresented. Third, cross-domain validation is needed to assess the generalisability of the framework across different subject areas and item types. Fourth, future work may explore hybrid modelling approaches that combine transformer-based architectures with additional linguistic or pedagogical features to improve the classification of lower-order cognitive levels.

Finally, advancing this line of research requires moving beyond standalone classification toward full system integration, including simulation and live deployment within adaptive testing platforms. Such efforts will be critical in establishing the practical value of cognitively informed machine learning approaches in educational assessment and in supporting the development of next-generation adaptive testing systems that more accurately reflect the multidimensional nature of learning.

REFERENCES

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2021). A systematic review of data-driven approaches to item difficulty prediction. *International Conference on Artificial Intelligence in Education*, 29–41.

An, X., & Yung, Y.-F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS Institute Inc*, *10*(4), 364–2014.

Asthana, P., Mishra, S., & Hazela, B. (2024). Text Identification for Questions Generation According to Bloom's Taxonomy Using Natural Language Processing. In *Machine Learning in Educational Sciences: Approaches, Applications and Advances* (pp. 335–357). Springer. https://doi.org/https://doi.org/10.1007/978-981-99-9379-6_16

Banda, S., Phiri, F., Kaale, J., Banda, A. M., Mpolomoka, D. L., Chikopela, R., & Mushibwe, C. (2023). Application of bloom's taxonomy in categorization of cognitive process development in colleges. *Journal of Education and Practice*, *14*(4), 6–13.

Besar, M. N. A., Abd Aziz, K. H., & Halim, H. A. (2025). The Validity of Multiple-True-False and One-Best-Answer in the Final Professional Undergraduate Medical Examination. *Education in Medicine Journal*, *17*(2), 5–21.

Bhopale, A. P., & Tiwari, A. (2024). Transformer based contextual text representation framework for intelligent information retrieval. *Expert Systems with Applications*, *238*, 121629.

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142–151.

Chandio, M. T., Pandhiani, S. M., & Iqbal, R. (2016). Bloom's Taxonomy: Improving Assessment and Teaching-Learning Process. *Journal of Education and Educational Development*, *3*(2), 203–221.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., & Gonçalves, M. A. (2023). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM Computing Surveys*, *55*(13s), 1–52.

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. *Computational Intelligence and Neuroscience*, *2022*(1), 1883698.

Du, X., Yang, J., Hung, J.-L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*, *48*(4), 225–236.

Forehand, M. (2010). Bloom's taxonomy. *Emerging Perspectives on Learning, Teaching, and Technology*, *41*(4), 47–56. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e277af328821afe28b0c9f506319ce6bf7228e5e

Gavhane, J. M., & Pagare, R. (2025). Revolutionizing Academic Evaluation: Bloom's Taxonomy Meets Deep Learning and NLP. *2025 IEEE Global Engineering Education Conference (EDUCON)*, 1–10.

ISSN: 2790-4407 www. jocatia.acata.org

Lu, C., & Cutumisu, M. (2021). Integrating Deep Learning into an Automated Feedback Generation System for Automated Essay Scoring. *International Educational Data Mining Society*.

Madhavi, A., Nagesh, A., & Govardhan, A. (2023). Efficient Course Recommendation using Deep Transformer based Ensembled Attention Model. *EAI Endorsed Transactions on E-Learning*, *9*.

Meissner, R., Jenatschke, D., & Thor, A. (2020). Evaluation of approaches for automatic e-assessment item annotation with levels of Bloom's taxonomy. *International Symposium on Emerging Technologies for Education*, 57–69. https://doi.org/https://doi.org/10.1007/978-3-030-66906-5_6

Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, 267–281.

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. *International Conference on Artificial Intelligence in Education*, 469–481.

Weibelzahl, S., Paramythis, A., & Masthoff, J. (2020). Evaluation of adaptive systems. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 394–395.

Zaidi, N. L. B., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? *Academic Medicine*, *93*(6), 856–859.

Zhang, L., & VanLehn, K. (2016). How do machine-generated questions compare to human-generated questions? *Research and Practice in Technology Enhanced Learning*, *11*(1), 7.