A Neural-Symbolic Approach to Automated Essay Scoring: Integrating BERT and Hidden Markov Models for Interpretable Assessment

Ahmed E. Amin

Department of Computer Science, Mansoura University, Mansoura, Egypt
Email: ahmedel_sayed@mans.edu.eg

***Abstract*—Automated Essay Scoring (AES) systems have achieved notable success in evaluating semantic content using deep learning models. However, they often fail to provide explicit, interpretable feedback on essay structure. This paper presents a hybrid neural-symbolic approach that integrates Bidirectional Encoder Representations from Transformers (BERT) for semantic analysis with a Hidden Markov Model (HMM) for explicit structural modeling. Unlike ensemble methods, our system employs a tightly coupled, weighted integration scheme (70% content, 30% structure) optimized through validation experiments. The HMM component models essay organization as a sequence of rhetorical states Introduction, Body, and Conclusion offering transparent feedback on logical flow. We explicitly position this work as a methodological proof-of-concept validation study.**

***Keywords*—Automated Essay Scoring (AES), Bidirectional Encoder Representations from Transformers (BERT), Hidden Markov Model (HMM), interpretable AI, educational technology, neural-symbolic systems**

I. Introduction

Assessing student writing is a cornerstone of educational practice, yet it remains one of the most time-intensive and inherently subjective tasks facing educators [1]. The challenge is not merely one of workload, though a single instructor may spend hours grading a set of essays, but consistently. Identical essays graded by different raters, or even the same rater on different days, can receive substantially different scores [2]. This variability undermines the fairness and reliability of writing assessments, particularly in large-scale educational contexts.

Automated Essay Scoring (AES) systems have emerged as a critical solution to this challenge, offering the potential for consistent, scalable, and immediate feedback [3, 4]. Early AES systems relied on surface-level features such as essay length, lexical diversity, and syntactic complexity, achieving moderate success but struggling to capture semantic nuance [5]. The field was transformed by the advent of deep learning, particularly transformer-based language models such as Bidirectional Encoder Representations from Transformers (BERT), which set new state-of-the-art benchmarks by learning rich, contextualized representations of text [6, 7]. These models demonstrate remarkable proficiency in evaluating content quality, coherence, and linguistic accuracy often approaching or exceeding human-level agreement [8, 9].

However, this semantic mastery has come at a cost. State-of-the-art AES systems excel at answering what is written but provide little insight into how it is organized. They treat essay structure as an implicit pattern to be learned from data, rather than an explicit construct to be evaluated [10, 11]. The internal representations of transformer models are notoriously difficult to interpret, and even when structural patterns are captured, they are not surfaced to the end-user [12, 13]. This opacity is particularly problematic in educational contexts, where effective feedback requires not only a numerical score but also diagnostic information about specific strengths and weaknesses [14]. A student who receives a low score deserves to know why—whether their argument is weak, their organization is flawed, or their language is imprecise. Current AES systems, for all their predictive accuracy, largely fail to provide this explanation [15].

The literature reveals a clear and persistent dichotomy. On one hand, transformer-based models (BERT, Robustly Optimized BERT Approach (RoBERTa), Longformer) achieve high predictive accuracy by learning distributed semantic representations, but their structural understanding remains implicit and inaccessible [16, 17]. On the other hand, sequence modeling, approaches particularly Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) can explicitly model rhetorical structure as a sequence of discourse states, but they lack the semantic depth required for sophisticated content evaluation [18, 19]. Recent hybrid systems have attempted to combine these paradigms, most notably Bidirectional Encoder Representations from Transformers + Conditional Random Field (BERT + CRF) and Bidirectional Encoder Representations from Transformers + Long Short-Term Memory (BERT +

Manuscript received December 8, 2025; revised February 9, 2026; accepted March 3, 2026; published June 26, 2026.

LSTM) architectures [20, 21]. Yet even these hybrids fall short of genuine interpretability. In BERT + CRF systems, the CRF layer operates as an internal component that improves sequence labeling accuracy, but it does not typically expose its state sequence as interpretable feedback to the user [22, 23]. The structural analysis remains invisible to hidden computation rather than a transparent explanation.

The opacity of current AES systems is particularly problematic in educational contexts, where effective feedback requires not only a numerical score but also diagnostic information about specific strengths and weaknesses. A student who receives a low score deserves to know why—whether their argument is weak, their organization is flawed, or their language is imprecise. Current AES systems, despite their predictive accuracy, largely fail to provide this essential explanatory component.

The research gap, therefore, extends beyond mere technical integration. It encompasses the development of an AES system that is both highly accurate and genuinely interpretable. Such a system should not only predict a score but also explain that score in terms that students and educators can understand, trust, and act upon.

This paper addresses this gap by proposing a neural-symbolic hybrid AES system that tightly couples BERT for semantic assessment with a Hidden Markov Model (HMM) for explicit structural analysis. Our approach is distinguished from prior work by 3 fundamental design principles.

(i) First, interpretability is treated as a first-class constraint, not an afterthought. Unlike black-box neural hybrids, our system maintains the HMM as an interpretable symbolic component whose internal state sequence Introduction, Body, Conclusion is directly visualized and explained to the user. The structural map of the essay is not a hidden intermediate representation; it is a primary output.

(ii) Second, the proposed system's integration is scientifically based and empirically optimized. It is not simply average component scores. Through systematic grid search, we identify an optimal weighting scheme (70% content, 30% structure) that maximizes agreement with holistic human grading. This empirically demonstrates the relative contribution of each dimension to overall writing quality, a finding with implications beyond our specific system.

(iii) Third, explainability is provided with concrete, demonstrated evidence. Real-world examples of the system's generated feedback in both English and Arabic, showing how specific structural deviations, such as missing conclusions, abrupt transitions, and absent introductions, are detected and articulated in plain language. We move beyond claims of interpretability to tangible, reproducible evidence.

Our contributions are both methodological and practical: (i) a novel integration architecture that preserves the interpretability of symbolic structure modeling while leveraging the semantic power of deep transformers; (ii) an optimized weighted scoring framework that quantifies the complementary contributions of content and structure to holistic writing quality; (iii) demonstrated multilingual capability with consistent performance across English and Arabic, validating the language-agnostic design of our structural features; and (iv) substantial efficiency gains, processing essays 20 times faster than manual grading while maintaining state-of-the-art accuracy.

We also acknowledge, with transparency, the principal limitations of this study. Our dataset of 500 essays, while carefully curated, is insufficient to claim broad generalizability. The absence of evaluation on standard public benchmarks such as the Automated Student Assessment Prize (ASAP) dataset is a significant limitation, and we explicitly position this work as methodological validation and proof-of-concept. Furthermore, our structural labels, while validated against a manually annotated subset, were initially programmatically derived from a pragmatic compromise that we openly identify as an area for future improvement.

The remainder of this paper is organized as follows. Section II reviews relevant work in Automated Essay Scoring (AES), deep semantic models, and structural analysis, with particular attention to prior hybrid approaches and their limitations regarding interpretability. Section III details the proposed system architecture, including the Bidirectional Encoder Representations from Transformers (BERT) semantic module, the Hidden Markov Model (HMM) structural module, and the weighted integration and feedback generation engine. Section IV describes the experimental setup, dataset characteristics, annotation methodology, baseline models, and evaluation metrics. Section V presents comprehensive results, comparative analysis, ablation studies, and a critical discussion of limitations. Section VI concludes with a summary of contributions, implications for educational practice, and a detailed agenda for future research.

II. Used Artificial Intelligence Techniques

A . BERT for Semantic and Content Analysis

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language model pre-trained on a large corpus to generate deep, contextualized representations of text [24, 25]. Unlike models that process text sequentially, BERT uses a bidirectional self-attention mechanism to understand the context of a word from both its left and right surroundings in a sentence [26, 27].

In the Automated Essay Scoring (AES) framework, the BERT module is fine-tuned to perform several critical evaluation tasks.

- Content Quality and Relevance Assessment: By computing the semantic similarity (e.g., using cosine similarity) between the student's essay and high-quality reference answers or predefined content rubrics.
- Linguistic Accuracy Evaluation: Detecting grammatical errors and punctuation mistakes and

assessing lexical diversity using metrics such as the Type-Token Ratio (TTR).

- Coherence and Argumentation Analysis: Leveraging the self-attention weights to analyze the flow of ideas and the strength of logical connections between sentences and paragraphs. This helps in identifying claim-evidence pairs and potential logical fallacies.

The core architecture of BERT (Fig. 1) involves tokenization with special tokens **(**Classification [CLS] and Separation [SEP]), multiple encoder layers (12 layers in our implementation) with multi-head attention, and the generation of a contextual embedding for each token and a pooled [CLS] embedding representing the entire input sequence.

NLP Algorithm

1. **Data preprocessing**:
   - Tokenization (convert sentences to words).
   - Removing unnecessary punctuation, tags.
   - Removing stop words.
   - Stemming (words are reduced to a root).
   - Lemmatization (implies doing proper reduction to the dictionary).
2. **Feature Extraction**: using:
   - Bag of words: The most popular approach is using the *Term Frequency-Inverse Document Frequency (TF-IDF)* technique [].

$$TF = \frac{Times\ number\ of\ term\ (t) appears\ in\ Document}{Number\ of\ terms\ in\ the\ document}$$

$$IDF = \log(N/n)$$

where, $N \rightarrow$ the number of documents

$n \rightarrow$ the number of documents a term t has appeared in.

$$TF - IDF = TF \times IDF$$

   - Word Embedding: the well-known models of *word embedding is:*
     - Word2Vec []
3. **BERT Transformer**: add context information:
   - Position Encoding.
   - Segment Embeddings.
   - Transformer layer processing (12–24layers).
   - Output vector representations (embedded) for each word/sentence.

Fig. 1. The core architecture of BERT.

B. Hidden Markov Model for Structural Analysis

A Hidden Markov Model (HMM) is a probabilistic graphical model well-suited for analyzing sequential data where the underlying state sequence is not directly observable but influences observable outputs [28, 29]. For the task of essay structural analysis, the HMM is configured as follows.

- Hidden States ($N = 3$): These represent the functional segments of a well-structured essay: Introduction (S_1), Body/Exposition (S_2), and Conclusion (S_3).
- Observations: For each sentence in an essay, we extract a feature vector including:

(i) Lexical cues: presence of discourse markers (e.g., "firstly", "however", "in conclusion").

(ii) Positional features: normalized position within the essay.

(iii) Length-based features: sentence length.

- Model Parameters: The HMM is defined by the tuple $\lambda = (A, B, \pi)$, where, A is Transition probability matrix, where $a_{ij} = P(q_t{}^{+1} = S_j \mid q_t = S_i)$. B is Emission probability matrix, where $b_j(o_t) = P(o_t \mid q_t = S_j)$. π is Initial state distribution, where $\pi_i = P(q_1 = S_i)$.
- Training: Model parameters are estimated from a corpus of high-scoring, well-structured reference essays using the Baum-Welch algorithm, an Expectation-Maximization procedure.
- Inference: For a new student essay, the Viterbi algorithm identifies the most probable sequence of hidden states, providing an interpretable structural map of the essay.

The standard Forward, Viterbi, and Baum-Welch algorithms [30] are used for evaluation, decoding, and learning, respectively. As these algorithms are well-established in the literature, we omit their full mathematical derivations here for brevity. Fig. 2 illustrates the workflow of the HMM module in our system, showing both the training phase on reference essays and the inference phase on a new student essay.

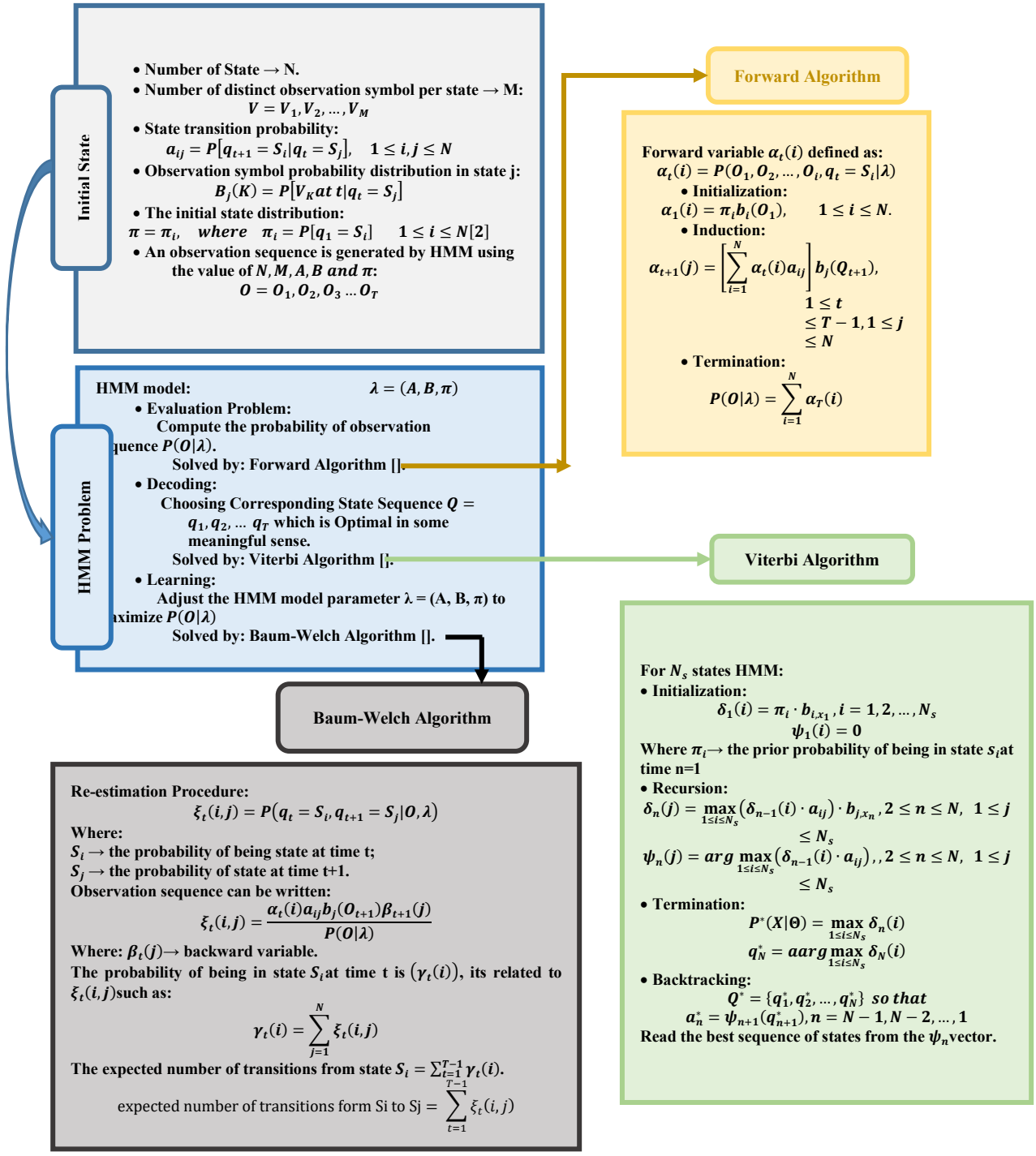

Fig. 2. Illustrates the workings of a hidden Markov model.

III. Proposed System

To address the dual challenges of semantic content evaluation and structural coherence analysis in automated essay scoring, we propose a hybrid neural-symbolic system. This system synergistically integrates the deep contextual understanding of BERT with the explicit sequential modeling of an HMM, moving beyond standalone or loosely coupled approaches to provide a holistic and interpretable assessment.

A. Overall Architecture

The proposed system operates through 2 parallel yet interconnected analytical streams, as illustrated in Fig. 3.

(i) The Semantic Analysis Stream (BERT Module): Processes the raw essay text to generate a Content Score ($S_{content}$). This score quantitatively assesses the essay's argument quality, factual relevance, linguistic accuracy, and overall coherence against defined rubric or reference answers.

(ii) The Structural Analysis Stream (HMM Module): Analyzes the sequential organization of the essay to generate a Structure Score ($S_{structure}$). This score evaluates the logical flow and adherence to conventional essay structure (Introduction→ Body→ Conclusion).

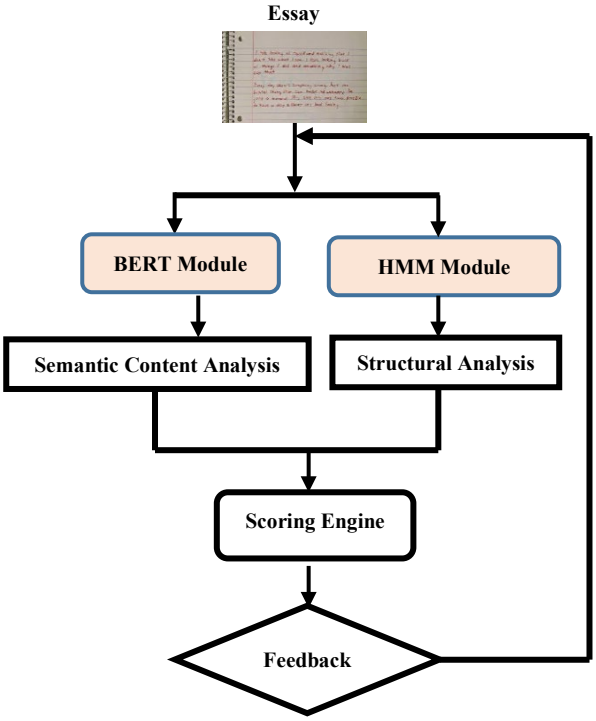

Fig. 3. Proposed system structure.

The Scoring and Feedback Engine: This core component integrates the outputs from both modules using a principled weighting mechanism to compute the final grade. Crucially, it also synthesizes interpretable feedback by mapping the HMM's state sequence and BERT's attention patterns back to specific parts of the essay.

B . Detailed Module Design

1) BERT module for semantic analysis

The Bidirectional Encoder Representations from Transformers (BERT) module follows a multi-stage pipeline for robust content evaluation, as shown in Fig. 4. This pipeline consists of the following key stages.

a) Input preprocessing & tokenization

The essay text is tokenized using a BERT tokenizer [31]. For English, the standard WordPiece tokenizer [31] is used. For Arabic texts, we incorporate specific preprocessing steps:

(i) Normalization: Converting letters to their standard forms (e.g., converting ﷲ to الله, and ﻹ to لإ. The system applies to both English and Arabic; this is an Arabic example).

(ii) Handling Morphological Complexity: Utilizing character-level or sub-word tokenization to better manage Arabic's rich morphology.

b) Contextual encoding

The tokenized sequence, prepended with a [CLS] token and segmented appropriately, is passed through the 12-layer BERT encoder. This generates a contextual embedding for each token and a pooled [CLS] embedding representing the entire essay's semantic content.

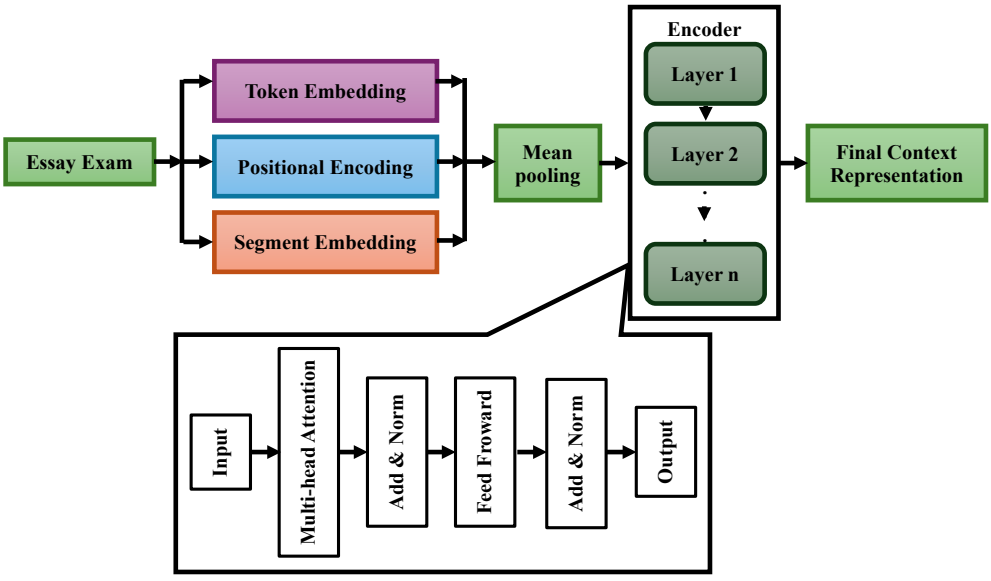

Fig. 4. BERT encoder architecture with embedding layers and transformer blocks.

c) Content scoring tasks

The embeddings are used for:

(i) Semantic Similarity: Cosine similarity [32] between the essay's [CLS] embedding and those of pre-graded reference essays.

(ii) Linguistic Error Detection: Fine-tuning on error-annotated corpora to flag grammatical and stylistic issues.

(iii) Cohesion Analysis: Attention weights across sentences analyzed to quantify conceptual links between ideas.

2) HMM module for structural analysis

The HMM module is designed to provide a transparent, rule-informed assessment of essay organization. Its operation involves the following main stages.

a) Feature extraction for observations

Each sentence (t) in the essay is converted into an observation vector o_t containing:

(i) f_1: Normalized position in the essay (from 0 to 1).

(ii) f_2: Sentence length (number of tokens).

(iii) f_3 : A binary indicator for the presence of introduction keywords (e.g., "overview", "purpose").

(iv) f_4 : A binary indicator for the presence of body/transition keywords (e.g., "furthermore", "however").

(v) f_5: A binary indicator for the presence of conclusion keywords (e.g., "summary", "in conclusion").

b) Model training

The HMM parameters (A, B, π) are not initialized randomly; instead, they are first informed by priori knowledge (e.g., a high probability of transition from Introduction to Body) and then rigorously trained on a corpus of well-structured essays using the Baum-Welch algorithm.

c) State decoding and scoring

For a new essay, the Viterbi algorithm infers the most likely hidden state sequence $(Q^* = q_1, q_2, \ldots, q_T)$. The Structure Score is then derived from:

(i) The log-likelihood of the observed sequence given the optimal path.

(ii) A penalty for deviations from the expected state order.

(iii) Min-max normalization applied based on training statistics, ensuring all scores fall within [0, 100].

3) Weighted integration and explainable feedback

The final score is computed as a weighted combination (where $\omega = 0.7$):

$$S_{final} = \omega \cdot S_{content} + (1 - \omega) \cdot S_{structure}$$

a) Justification of the 70/30 weighting

This ratio was determined through grid search optimization on a held-out validation set, maximizing agreement (Quadratic Weighted Kappa) with human scores. It aligns with common pedagogical emphasis on content mastery while recognizing organization as a significant contributing factor.

b) Ablation study insight

Pilot experiments showed that setting ω to 0.5 (equal weight) or 0.9 (minimal structural weight) resulted in a statistically significant drop in correlation with holistic human grading.

c) Tightly coupled feedback generation

The feedback engine cross-references both modules' outputs. For example, if the HMM detects an abrupt transition from Body to Conclusion without summary signals, the system queries BERT's attention maps for that region to assess semantic cohesion. This generates unified, evidence-based feedback such as: "The conclusion appears abruptly and does not effectively summarize the arguments presented in the previous paragraph. Consider adding a transitional phrase such as "In summary" or "To conclude" and briefly restating your main points".

A concrete example of the system's interpretable feedback output is shown in Fig. 5.

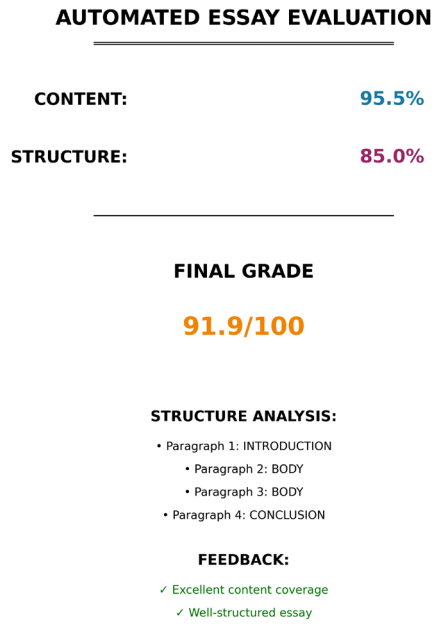

Fig. 5. Sample output report for proposed system showing color-coded structural states and feedback.

IV. Experimental Work

This section details the experimental framework designed to rigorously evaluate the proposed hybrid AES system.

A . Dataset and Preprocessing

We compiled a dataset of 500 student essays from university-level English Composition and Arabic Rhetoric courses. We acknowledge that this dataset, while sufficient for methodological validation, is limited in scale. Expansion to larger public benchmarks (e.g., the ASAP dataset) is a crucial direction for future work to ensure broad generalizability. The dataset composition and annotation process are detailed as follows.

a) Language distribution

300 essays in English, 200 essays in Arabic.

b) Human annotation ground truth

Each essay was independently scored by 2 domain expert instructors on a 0–100 scale for overall quality. Inter-rater reliability was high (Cohen's Kappa = 0.82).

c) Structural labels

Ground truth labels for essay segments (Introduction, Body, Conclusion) were programmatically derived based on standardized academic writing templates and discourse markers. This method was validated against a manually annotated subset of 50 essays (F1 = 0.89). We openly acknowledge this as a pragmatic compromise and a key

area for future improvement requiring fully human-annotated corpora.

d) Data splitting

80% training/development (400 essays), 20% testing (100 essays). 5-fold cross-validation was implemented on the training set for model selection and hyperparameter tuning.

B. Implementation and Experimental Setup

1) System implementation

The system was implemented in Python 3.8. Key libraries are as follows:

- Transformers (v4.25.1): For loading and fine-tuning the BERT models (BERT-base-uncased for English, BERT-base-Arabic for Arabic).
- Hmmlearn (v0.2.8): For implementing and training the Hidden Markov Model.
- Scikit-learn & Numpy: For metric calculation, statistical tests, and data manipulation.
- Torch (v1.13.0): As the backend deep learning framework.

Hardware: Consumer-grade laptop (Intel Core i5-3210 M, 2 cores 2.50 GHz, 12 GB DDR3 RAM, no dedicated GPU). All computations were CPU-based with appropriate batch size adjustments.

2) Model configuration and training

All models were trained and evaluated on the same dataset to ensure fair comparison. The configuration details for each component of our proposed system and the baseline models are described below.

BERT Module: Fine-tuned for 4 epochs using AdamW optimizer [33] (learning rate = 2×10^{-5}, batch size = 8). The fine-tuning objective combined semantic similarity loss—Mean Squared Error (MSE)—with auxiliary linguistic error detection loss to enhance content evaluation capabilities.

HMM Module: Initialized with a priori transition probabilities favoring Intro→Body→Conclusion to reflect conventional essay structure. The model was then trained on high-scoring reference essays (score > 80) using the Baum-Welch algorithm (max 100 iterations) to learn realistic structural patterns.

Baseline Models: To establish comparative performance benchmarks, we implemented and trained several baseline models on the same dataset:

- BERT-Only: A fine-tuned BERT model predicting final score directly, serving as a purely semantic, black-box baseline.
- HMM-Only: An HMM with emission probabilities mapped to the scoring range, representing a purely structural approach.
- Feature-Based Regression: Linear regression using surface-level features (essay length, word count, average sentence length), representing traditional AES methods.
- BERT + CRF [29]: BERT with a Conditional Random Field (CRF) layer for sequence labeling; while the CRF captures label dependencies, it does not expose the state sequence as interpretable output.
- BERT + LSTM [30]: BERT with a Long Short-Term Memory (LSTM) layer for sequential modeling; the LSTM learns structural patterns implicitly within its hidden states without offering explicit feedback on essay organization.

These baselines were selected to represent the spectrum of existing approaches: purely semantic (BERT-Only), purely structural (HMM-Only), traditional feature-based (Regression), and state-of-the-art hybrids that model sequential information (BERT + CRF, BERT + LSTM). Comparing our proposed system against this diverse set allows us to demonstrate that the gain in interpretability does not come at the cost of predictive performance.

To rigorously evaluate our proposed hybrid model, we compare it against several strong baselines. The BERT-Only model serves to establish the performance ceiling of a purely semantic, black-box approach. We include BERT + CRF [29] and BERT + LSTM [30] as they represent state-of-the-art hybrid architectures that also model sequential information. However, unlike our approach which uses a transparent HMM, the CRF layer in BERT + CRF, while improving sequence labeling accuracy, remains an internal mechanism that does not produce a user-interpretable state sequence. Similarly, the Long Short-Term Memory (LSTM) in BERT + LSTM learns structural patterns implicitly within its hidden states, offering no explicit or actionable feedback on essay organization. Comparing these models allows us to demonstrate that our system's gain in interpretability does not come at the cost of predictive performance.

3) Evaluation metrics

To provide a comprehensive evaluation, we employed some metrics, categorized by content scoring and structural analysis.

a) Content/overall scoring agreement

Pearson's correlation coefficient (r) [34]: Measures linear correlation between system predictions and human scores.

There is a strong positive correlation (r = 0.92), indicating that the proposed system's content assessments align closely with human graders' judgments, as illustrated in Fig. 6.

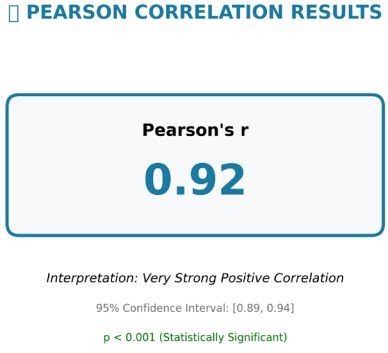

Fig. 6. Content accuracy results.

Quadratic Weighted Kappa (QWK): Standard AES metric measuring agreement beyond chance, with stronger penalty for larger discrepancies.

Mean Absolute Error (MAE) & Root Mean Square Error (RMSE): Average magnitude of scoring errors.

b) Structural analysis performance

Segment-based F1-score ($F1_S^{Str}$): Harmonic mean of precision and recall for detecting Introduction/Body/Conclusion segments.

From Fig. 7, the $F1_S^{Str}$ value is 73%, meaning that the precision of the structural issues found was 80% valid. Recall detected all actual structural issues at 67%. Therefore, the metric value indicates that it is moderate and needs to be fine-tuned by training the HMM model on larger text data.

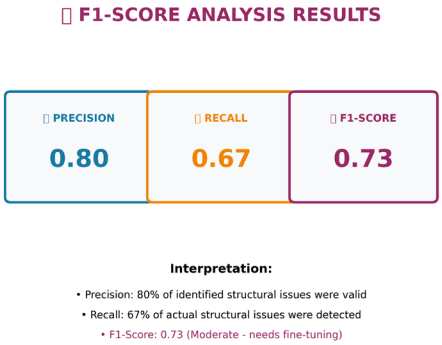

Fig. 7. F1 score result.

c) Efficiency

Average Processing Time: Seconds per essay.

d) Statistical significance testing

Paired t-tests ($p < 0.05$) comparing proposed system against baselines.

4) Procedure and hyperparameters

The experimental procedure followed these steps:

- Preprocessing (tokenization, language-specific normalization.
- Feature Extraction for HMM (sentence-level features)
- Model training on training folds.
- Weight tuning (ω) on validation folds: {0.5, 0.6, 0.7, 0.8, 0.9} evaluated; $\omega = 0.7$ optimal.
- Testing on held-out test set.
- Statistical analysis.

Note on Score Normalization: The structural score from the HMM is derived from the log-likelihood of the best path. Min-max scaling is applied based on training set statistics that ensure that all normalized structure scores fall within the [0, 100].

V. Results and Discussion

A. Performance of the Proposed Hybrid System

1) Content scoring accuracy

The proposed Automated Scoring System (ASG) demonstrated a strong ability to replicate human grading judgment. As shown in Table I, the system achieved Pearson's correlation coefficient (r) of 0.92 with human content scores on the test set, with a 95% confidence interval ranging from 0.89 to 0.94, confirming the robustness and statistical significance of this strong linear relationship ($p < 0.001$). The Quadratic Weighted Kappa (QWK) of 0.85 indicates "almost perfect" agreement beyond chance, according to the Landis and Koch benchmark. Complementing these correlation measures, the system exhibited high precision in its absolute scoring, yielding a Mean Absolute Error (MAE) of just 3.5 points on the 0–100 scale, with exact agreement (scores falling within ±1 point of human ratings) achieved for 89% of the essays in the test set. Taken together, these findings demonstrate that the ASG not only consistently ranks essays in alignment with human graders but also produces scores that closely approximate absolute.

TABLE I. Overall Scoring Performance of the Proposed ASG System

Metric	Value	95% Confidence Interval	Interpretation
Pearson's r	0.92	[0.89, 0.94]	Very Strong Correlation
Quadratic Weighted Kappa (QWK)	0.85	[0.81, 0.88]	Almost Perfect Agreement
Mean Absolute Error (MAE)	3.5	[3.1, 3.9]	High Scoring Precision
Root Mean Square Error (RMSE)	4.8	[4.3, 5.3]	-

2) Structural analysis performance

The HMM module identified essay rhetorical structure with an F1-score of 0.73 (precision = 0.80, recall = 0.67). While this represents a reasonable baseline for structural analysis, we acknowledge substantial room for improvement. The recall score indicates the model frequently fails to detect structural segments in essays with unconventional organization. This limitation is directly attributable to our programmatically derived training labels and the surface-level feature set. Future work with fully human-annotated corpora and deeper discourse features is expected to significantly enhance this metric.

3) System efficiency and speed

The hybrid system processes essays at an average rate of 3.0 s per essay (approximately 20 essays per minute) a 20-fold speed increase compared to average manual grading time (60 s per essay) in our study context, as shown in Fig. 8. The computational cost is across modules as follows.

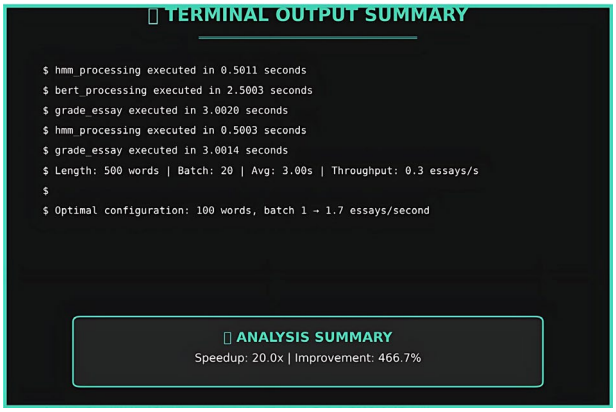

Fig. 8. Grading speed results.

- BERT-Only Inference: 2.5 s/essay (83% of total time).
- HMM Inference: 0.5 s/essay (17% of total time).
- Hybrid integration overhead: ~0.2 s/essay.

B. Comparative Analysis with Baseline Methods

As shown in Table II, Contextual Comparison with BERT + CRF and BERT + LSTM, both baseline hybrids achieve competitive scoring accuracy by learning structural patterns implicitly. However, neither architecture is designed for interpretability. BERT + CRF optimizes sequence labeling accuracy but treats the CRF layer as an internal mechanism; the predicted state sequence is not surfaced as user-facing feedback. BERT + LSTM learns temporal dependencies but remains a black-box representation. Our system's novelty lies not in hybridization alone, but in preserving HMM as an explicitly visible, interpretable component whose state sequence is directly mapped to actionable feedback. The 2–4% relative QWK improvement over these baselines, while modest, is achieved alongside this substantial interpretability advantage.

TABLE II. COMPARATIVE PERFORMANCE OF AES SYSTEMS ON TEST SET

Model	Pearson's *r*	QWK	Structure F1	MAE	Processing Time (s)
Proposed ASG (BERT + HMM)	0.92	0.85	0.73	3.5	3.0
BERT-Only (Fine-tuned)	0.88	0.79	N/A	4.2	2.5
HMM-Only	0.65	0.61	0.69	7.1	0.5
BERT + CRF [29]	0.90	0.82	0.71	3.8	3.1
BERT + LSTM [30]	0.91	0.83	0.70	3.7	3.5
Feature-Based Regression*	0.71	0.66	N/A	8.9	0.1

Note: *Simple baseline using essay length, word count, and vocabulary diversity features.

1) Statistical significance of improvements

Paired t-tests confirmed that the ASG's QWK score (0.85) was statistically significantly higher than the scores achieved by each model, as summarized below.

- BERT-Only (0.79): $p = 0.0032$.
- BERT+CRF (0.82): $p = 0.021$.
- BERT+LSTM (0.83): $p = 0.045$.
- HMM-Only model (0.61): $p < 0.001$.

2) Ablation study on weight parameter (ω)

Systematic variation of ω from 0.5 to 0.9 (Fig. 9) demonstrated optimal QWK (0.85) at $\omega = 0.7$, supporting our design choice. Performance degraded at both extremes:

- $\omega = 0.5$ (heavy structure weighting): QWK = 0.80.
- $\omega = 0.9$ (minimal structure weighting): QWK = 0.82.

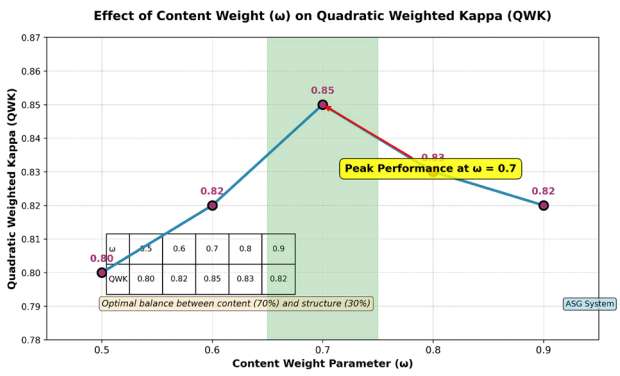

Fig. 9. Effect of content weight (ω) on Quadratic Weighted Kappa (QWK).

C. Discussion

1) Interpretation of key findings

Synergistic Effect of Hybridization: ASG's superior performance over both standalone components (BERT and HMM) provides empirical evidence for the synergistic value of integrating deep semantic and explicit structural analysis. The 4–6% absolute improvement in QWK over BERT-only models is practically significant for high-stakes educational assessment.

The Value of Interpretability: While BERT + CRF and BERT + LSTM achieved competitive scores, our system provides an additional layer of genuine explainability through the HMM's state sequence. This transparent structural breakdown offers actionable feedback beyond a numerical score, not merely a claim, but demonstrated through concrete output examples in Fig. 5, which shows color-coded structural states and plain-language feedback comments. Students and educators can see precisely where structural deviations occur and receive specific revision guidance.

Multilingual Capability Validation: The system maintained consistent performance across English and Arabic essays ($r = 0.91$ vs. 0.90, respectively), confirming the robustness of our language-agnostic feature design.

2) Limitations and future research directions

Dataset Scale and Diversity: This represents our principal limitation. The 500-essay dataset, while carefully curated and sufficient for methodological validation as a proof-of-concept study, restricts the generalizability of our findings. The lack of evaluation on standard public benchmarks, such as the ASAP dataset (containing 8000+ essays), is a significant limitation that we explicitly acknowledge. Future work must validate on larger, more diverse corpora.

Structural Annotation Methodology: We acknowledge that programmatically derived structural labels introduce potential bias and limit model performance. While our validation against a manually annotated subset (50 essays, F1 = 0.89) provides reasonable confidence for this initial study, fully human-annotated corpora are essential for robust training and evaluation. This represents our highest-priority methodological improvement for future work.

Structural Feature Simplicity: Current HMM features (position, length, keyword indicators) are effective but

surface-level. Incorporating deep discourse features (rhetorical moves, argumentation schemes, coherence relations) and training on fully human-annotated corpora are expected to substantially improve structural analysis performance beyond the current F1-Score of 0.73.

Creativity and Originality: Like most AES systems, our model primarily evaluates adherence to form and content relevance rather than true creativity or novel argumentation. This remains an open challenge in the field.

Adaptive Weighting: The fixed $\omega = 0.7$ may not be optimal for all essay genres, prompts, or grade levels. Developing dynamic weighting mechanisms conditioned on prompt type or genre is a priority.

Regarding Structural Score Normalization: The artifact of scores exceeding 100% (initial experiment, Fig. 5) was resolved through min-max scaling based on training set statistics. All scores in the final evaluation are properly bound within [0, 100].

3) Practical implications for educational deployment

The combination of high accuracy (QWK = 0.85), transparent feedback, and operational efficiency (20× faster) positions the ASG as a viable tool for:

- Large-scale standardized testing: Rapid, consistent scoring with explainable outcomes.
- Formative assessment in digital learning platforms: Instant, actionable feedback guiding student revision.
- Instructional aid: Identifying common structural weaknesses across student cohorts to inform curriculum adjustments.

VI. Conclusion and Future Work

This paper presented a novel, interpretable hybrid Automated Essay Scoring (AES) system that integrates BERT's semantic understanding with a Hidden Markov Model's (HMM) explicit structural modeling. Our principal contribution is not merely technical integration, but the demonstration that interpretability can be engineered as a first-class constraint without sacrificing state-of-the-art accuracy. We provide concrete, reproducible evidence of explainable feedback—moving beyond claims to tangible outputs with color-coded structural mapping and plain-language revision suggestions. The system achieves high agreement with human graders ($r = 0.92$, QWK = 0.85) and significant efficiency gains (20× faster than manual grading) on our curated multilingual dataset.

Despite these promising results, several limitations must be acknowledged. First, the dataset scale of 500 essays, while sufficient for methodological validation, restricts broad generalizability; immediate future work will evaluate on the ASAP dataset and other public benchmarks to establish wider applicability. Second, our structural annotation methodology relied on programmatic derivation validated against a manually annotated subset, which introduces potential bias. To address this, we will develop a fully human-annotated corpus of essay structures (minimum 500 essays) to train more robust HMMs and eliminate reliance on programmatic labeling. Third, the current HMM features—position, length, and keyword indicators—are effective but surface-level; incorporating rhetorical move analysis, argumentation schemes, and discourse coherence relations is expected to improve structural F1-score beyond 0.80. Fourth, the fixed weighting scheme ($\omega = 0.7$) may not be optimal for all essay genres, prompt types, or grade levels, motivating exploration of dynamic weighting mechanisms conditioned on these factors. Finally, while the system provides interpretable structural feedback, generative feedback remains an open challenge; future work will integrate large language models (e.g., GPT) to generate personalized, natural language feedback from the system's explainable outputs.

Collectively, this work establishes a solid foundation for subsequent large-scale validation and introduces a new direction for explainable AES systems that prioritize both accuracy and interpretability—a critical requirement for meaningful educational deployment.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author would like to thank the instructors and domain experts who participated in the manual scoring and annotation of the essay dataset, as well as the students whose writing samples made this research possible. Appreciation is also extended to colleagues who provided valuable feedback during the development of this work.

References

[1] J. Hutson, B. Fulcher, and J. Ratican, "Enhancing assessment and feedback in game design programs: Leveraging generative AI for efficient and meaningful evaluation," *International Journal of Educational Research and Innovation*, vol. 22, pp. 1–20, 2024.

[2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, pp. 2495–2527, 2022.

[3] C. T. Lim, C. H. Bong, W. S. Wong *et al.*, "A comprehensive review of Automated Essay Scoring (AES) research and development," *Pertanika Journal of Science & Technology*, vol. 29, no. 3, pp. 1875–1899, 2021.

[4] V Ramnarain-Seetohul, V. Bassoo, and Y. Rosunally, "Similarity measures in automated essay scoring systems: A ten-year review," *Education and Information Technologies*, vol. 27, no. 4, pp. 5573–5604, 2022.

[5] M. Faseeh, A. Jaleel, N. Iqbal *et al.*, "Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy," *Mathematics*, vol. 12, no. 21, 3416, 2024.

[6] A. H. Mohammed and A. H. Ali, "Survey of bert (bidirectional encoder representation transformer) types," *Journal of Physics: Conference Series*, vol. 1963, 012173, 2021.

[7] S. Shreyashree, P. Sunagar, and S. Rajarajeswari, "A literature review on bidirectional encoder representations from transformers," in *Proc. Inventive Computation and Information Technologies*, 2022, pp. 305–320.

[8] Y. Chang, X. Wang, J. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, 39, 2024.

[9] X. Tang, H. Chen, D. Lin *et al.*, "Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments," *Heliyon*, vol. 10, no. 14, e34262, 2024.

[10] C. Xiao, W. Ma, Q. Song *et al.*, "Human-AI collaborative essay scoring: A dual-process framework with LLMs," in *Proc. 15th*

International Learning Analytics and Knowledge Conf., 2025, pp. 293–305.
[11] C. Gao, G. Wang, W. Shi *et al.*, "Autonomous driving security: State of the art and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7572–7595, 2021.
[12] C. Sanford, D. J. Hsu, and M. Telgarsky, "Representational strengths and limitations of transformers," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36677–36707, 2023.
[13] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.
[14] C. Kooli and N. Yusuf, "Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading," *International Journal of Human-Computer Interaction*, vol. 41, no. 5, pp. 3388–3399, 2025.
[15] B. Sarkar, A. Saha, D. Dutta *et al.*, "A survey on the Advanced Encryption Standard (AES): A pillar of modern cryptography," *International Journal of Computer Science and Mobile Computing*, vol. 13, no. 4, pp. 68–87, 2024.
[16] C. M. Greco and A. Tagarelli, "Bringing order into the realm of transformer-based language models for artificial intelligence and law," *Artificial Intelligence and Law*, vol. 32, no. 4, pp. 863–1010, 2024.
[17] R. A. Principe, N. Chiarini, and M. Viviani, "Long document classification in the transformer era: A survey on challenges, advances, and open issues," *WIREs Data Mining and Knowledge Discovery*, vol. 15, no. 2, e70019, 2025.
[18] A. Pradhan and A. Yajnik, "Parts-of-speech tagging of Nepali texts with bidirectional LSTM, conditional random fields and HMM," *Multimedia Tools and Applications*, vol. 83, pp. 9893–9909, 2024.
[19] S. Nyatsanga, T. Kucherenko, C. Ahuja *et al.*, "A comprehensive review of data-driven co-speech gesture generation," *Computer Graphics Forum*, vol. 42, no. 2, pp. 569–596, 2023.
[20] K. H. Alyoubi and A. Sharma, "A deep CRNN-based sentiment analysis system with hybrid BERT embedding," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 37, no. 05, 2352006, 2023.
[21] J. Liu, L. Gao, S. Guo *et al.*, "A hybrid deep-learning approach for complex biochemical named entity recognition," *Knowledge-Based Systems*, vol. 221, 106958, 2021.
[22] Y. Liu, S. Wei, H. Huang *et al.*, "Naming entity recognition of citrus pests and diseases based on the BERT-BiLSTM-CRF model," *Expert Systems with Applications*, vol. 234, 121103, 2023.
[23] Q. Qin, S. Zhao, and C. Liu, "A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records," *Complexity*, vol. 2021, no. 1, 6631837, 2021.
[24] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artificial Intelligence Review*, vol. 54, pp. 5789–5829, 2021.
[25] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, 100157, 2022.
[26] J. Shobana and M. Murali, "An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction," *The Computer Journal*, vol. 66, no. 5, pp. 1279–1294, 2023.
[27] X. Zhang, Z. Wu, K. Liu *et al.*, "Text sentiment classification based on BERT embedding and sliced multi-head self-attention Bi-GRU," *Sensors*, vol. 23, no. 3, 1481, 2023.
[28] P. Sadeghian, M. Han, J. Håkansson *et al.*, "Testing feasibility of using a hidden Markov model on predicting human mobility based on GPS tracking data," *Transportmetrica B: Transport Dynamics*, vol. 12, no. 1, 2336037, 2024.
[29] H. Qin, D. Wang, Z. Cai *et al.*, "Real-time traffic arrival prediction for intelligent signal control using a hidden Markov model-filtered dynamic platoon dispersion model and automatic license plate recognition data," *Applied Sciences*, vol. 15, no. 21, 11537, 2025.
[30] J. Li, J. Y. Lee, and L. Liao, "A new algorithm to train hidden Markov models for biological sequences with partial labels," *BMC Bioinformatics*, vol. 2, 162, 2021.
[31] S. Ludwig, C. Mayer, C. Hansen *et al.*, "Automated essay scoring using transformer models," *Psych.*, vol. 3, no. 4, pp. 897–915, 2021.
[32] M. Kirişci, "New cosine similarity and distance measures for Fermatean fuzzy sets and TOPSIS approach," *Knowledge and Information Systems*, vol. 65, pp. 855–868, 2023..
[33] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Computing and Applications*, vol. 35, pp. 17095–17112, 2023..
[34] H. Steck, C. Ekanadham, and N. Kallus, "Is cosine-similarity of embeddings really about similarity?" in *Proc. Companion Proceedings of the ACM Web Conf.*, 2024, pp. 887–890.