© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718458

CHATUR: A Semantic-Based Social Learning and Conceptual Assessment Framework for Automated Descriptive Answer Evaluation

HRUSHIKESH SHASHIKANT GAWADE[1], DR. NETRAJA C. MULAY [2]
[1,2] *Master of Computer Application (MCA) Progressive Education Society's Modern College of Engineering*

Abstract- ***Traditional automated assessment systems used in e-learning platforms mainly depend on keyword matching and basic textual similarity methods. While these methods can detect lexical similarities and repeated expressions, they often struggle to interpret contextual meaning and accurately measure conceptual understanding. This issue becomes more prominent in modern social learning environments, where learners actively participate in generating, sharing, and evaluating knowledge collaboratively. Current assessment systems also struggle to provide detailed concept-level analysis, identify misconceptions accurately, classify responses according to cognitive complexity, and generate transparent feedback for learners. In addition, many advanced AI-driven grading approaches require substantial computational resources, which limits their practical applicability in scalable or resource-constrained educational platforms. To address these limitations, this research proposes the CHATUR framework designed for integration within a collaborative social learning and assessment platform similar to Reddit and X. In this environment, users can both contribute educational content and learn from shared community knowledge. The proposed system employs lightweight transformer-based semantic embedding techniques to evaluate descriptive answers efficiently in a dynamic user-driven setting. It performs contextual similarity analysis between student responses and reference answers, extracts key concepts to examine concept coverage, and identifies misconceptions using rule-based analysis methods. The framework also applies multi-dimensional scoring to assess partial correctness and uses cognitive-level classification to evaluate the depth of understanding demonstrated in responses. Furthermore, CHATUR incorporates explainable AI techniques to provide users with clear and meaningful feedback regarding their performance. Since the framework is optimized for CPU-based execution, it remains accessible, scalable, and suitable for large-scale deployment in low-resource educational environments. By combining AI-powered assessment with collaborative knowledge sharing, the proposed system aims to improve user engagement while delivering more accurate and insightful evaluation of conceptual understanding.***

Keywords— Semantic Assessment, Sentence-BERT, Descriptive Answer Evaluation, Explainable AI (XAI), Natural Language Processing (NLP), Conceptual Understanding, Semantic Similarity, Social Learning Platform, Automated Assessment System, Transformer-Based Learning, Educational Technology, AI-Based Evaluation.

I. INTRODUCTION

The expansion of digital education has transformed conventional learning by introducing platforms that support communication, collaboration, and knowledge sharing among users. Social learning applications similar to Reddit and X encourage learners to participate actively by posting content, discussing ideas, and evaluating shared information. Although these platforms improve learner engagement, the automated assessment methods used in many systems still rely heavily on keyword comparison and basic text-matching techniques. Such approaches often fail to understand the actual meaning of student responses or measure conceptual understanding accurately.

Another major limitation of existing assessment systems is their inability to analyse answers at the concept level, identify learner misconceptions, and provide meaningful feedback that enables students to strengthen their understanding. In addition, several AI-based evaluation models depend on high-end computational resources, which makes large-scale deployment difficult in low-resource educational environments.To address these issues, this research

introduces the Chatur, integrated into a collaborative social learning platform. The proposed framework uses transformer-based language models such as BERT

II. LITERATURE REVIEW

The development of automated assessment systems has progressively evolved from conventional keyword-oriented evaluation methods to more intelligent semantic-based learning approaches. Earlier assessment techniques primarily depended on keyword matching, lexical comparison, and predefined rule-based grading mechanisms. While these approaches were computationally efficient and relatively straightforward, they were unable to interpret the actual contextual meaning of descriptive answers or accurately evaluate conceptual understanding. Consequently, students expressing correct ideas using alternative wording or different sentence structures were often evaluated inaccurately.

With the advancement of Natural Language Processing and transformer-based deep learning models, automated educational assessment has become more context-aware and semantically focused. Modern AI-driven systems are capable of analysing relationships between words, understanding contextual information, and evaluating answers based on semantic relevance rather than exact keyword overlap. These developments have facilitated the creation of intelligent assessment frameworks that support semantic similarity analysis, concept extraction, explainable feedback generation, and more adaptive evaluation techniques for interactive digital learning environments.

1. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks — Reimers, N. & Gurevych, I. (2019)
This foundational work introduced Sentence-BERT (SBERT), an extension of the BERT architecture that employs siamese and triplet network structures to generate semantically meaningful sentence embeddings. Unlike conventional BERT, SBERT produces fixed-length vector representations that can be efficiently compared using cosine similarity. This significantly improves performance in semantic textual similarity tasks and makes the model appropriate for applications such as automated grading and information retrieval. SBERT has since become a core component in many modern semantic-similarity-based evaluation systems.

2. The Eras and Trends of Automatic Short Answer Grading — Burrows, S., Gurevych, I. & Stein, B. (2015)
This survey examined the historical evolution of Automatic Short Answer Grading (ASAG) systems, beginning with lexical matching approaches and progressing toward latent semantic analysis and neural-network-based methods. The authors categorized grading approaches into knowledge-based, corpus-based, and machine-learning-based techniques while critically discussing their limitations in handling paraphrasing, conceptual understanding, and partial credit assignment. The study concluded that keyword-based and surface-level similarity approaches are inadequate for accurately evaluating conceptual correctness in student responses.

3. Automatic Classification of Learning Objectives Based on Bloom's Taxonomy — Li, Y. et al. (2022)
This study collected over 21,000 learning objectives from an Australian university and applied several machine learning classifiers, including Naive Bayes, Support Vector Machines, Random Forest, and BERT-based models, to classify learning objectives according to Bloom’s Taxonomy cognitive levels. The BERT-based approaches achieved the best performance, reaching Cohen’s Kappa values up to 0.93 and F1-scores up to 0.95. However, the work focused only on predefined learning objectives and did not address the evaluation of open-ended student responses or descriptive answer grading.

4. Automated Grading using Natural Language Processing and Semantic Analysis — Ayaan, A. & Ng, K.W. (2025)
This research proposed an NLP-based automated grading framework that combined multiple similarity measures, including Jaccard similarity, edit distance, cosine similarity, and normalized word count, along with semantic analysis using TensorFlow’s Universal Sentence Encoder. The hybrid approach improved grading accuracy compared to purely lexical methods by integrating both surface-level and semantic analysis. Nevertheless, the system lacked advanced

capabilities such as concept extraction, misconception detection, cognitive-level classification, and explainable feedback generation.

5. Exam Questions Classification Based on Bloom's Taxonomy Cognitive Level Using Classifiers Combination — Abduljabbar, D.A. & Omar, N. (2015)
This study proposed a hybrid classification framework utilizing Decision Trees, K-Nearest Neighbors, and Naive Bayes algorithms to classify examination questions according to Bloom's Taxonomy cognitive domains. The findings demonstrated that combining multiple classifiers improved classification accuracy compared to individual models. However, the research focused exclusively on question classification rather than evaluating descriptive student answers or generating feedback based on conceptual understanding.

6. AI-Empowered Collaborative Assessment Study (2024)
This quasi-experimental study involving 135 college students proposed an AI-assisted assessment framework that improved collaborative
knowledge building, cognitive engagement, socially shared regulation behaviors, and overall group performance in online collaborative learning environments. Despite these positive outcomes, the framework relied heavily on instructor-designed content and structured collaborative settings instead of user-generated learning environments. Furthermore, the study did not incorporate semantic answer evaluation, misconception detection, or transparent explainable feedback mechanisms.

7. Developing Explainable AI Systems to Support Feedback for Students — EDM Doctoral Consortium (2024)
This research focused on the development of explainable AI systems powered by large language models to generate safe, reliable, and educationally meaningful feedback for students. The research also investigated differences between AI-generated and human-written feedback. Although the work highlighted explainable AI as a crucial requirement in educational assessment, it relied on resource-intensive large language models and did not address deployment challenges in resource-constrained environments.

III. RESEARCH GAP

1. Existing automated assessment systems fail to provide a unified framework that simultaneously integrates semantic similarity scoring, concept extraction, misconception detection, cognitive-level classification, and explainable feedback generation.
2. Most existing approaches fall into two extremes: either they rely on computationally expensive large language models, which are difficult to deploy at scale, or they use shallow lexical techniques such as keyword matching and TF-IDF, which lack deeper semantic understanding. Very few studies attempt to balance semantic intelligence with lightweight, CPU-efficient deployment suitable for scalable educational systems.
3. In addition, current automated grading systems have largely ignored social and collaborative learning environments in which user-generated content serves as both the learning resource and the assessment material. Existing studies primarily focus on instructor-prepared datasets and controlled academic settings.
4. Another significant research gap is the limited exploration of misconception detection in descriptive answers using lightweight NLP techniques. Most systems only evaluate whether an answer is correct or incorrect without identifying the underlying conceptual misunderstanding responsible for the error.
5. Finally, explainable feedback tailored specifically for learners remains insufficiently addressed in existing lightweight assessment systems. Most available approaches either provide generic feedback or rely on black-box AI models that lack transparency and interpretability.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718458

IV. FUNDAMENTALS OF BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMER (BERT)

A. Core Principle OF Bert

BERT is a transformer-based language model introduced by Google for complex Natural Language Processing (NLP) tasks. The model is designed to interpret the meaning of words by examining their contextual relationships within a sentence rather than treating words independently.

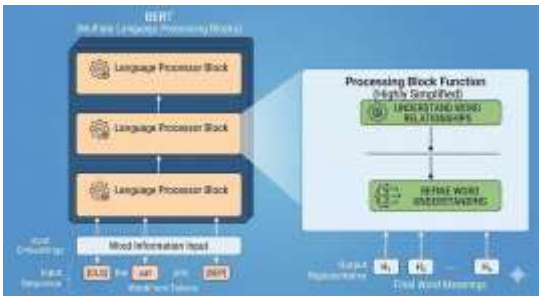
Unlike earlier language models that process text in only one direction, BERT examines both the preceding and following words simultaneously, allowing it to capture richer semantic relationships and contextual information. The architecture of BERT is built on the Transformer framework, which relies on self-attention mechanisms to identify how words in a sentence are contextually connected.

During processing, the model uses special tokens such as [CLS] and [SEP]. The [CLS] token stores the overall representation of the input sequence and is mainly used for classification-related tasks, whereas the [SEP] token helps distinguish separate sentences or segments within the input.

Before being applied to specific NLP tasks, BERT undergoes large-scale pre-training using two learning strategies. In Masked Language Modeling (MLM), certain words in a sentence are hidden, and the model learns to predict the masked words using contextual information. In Next Sentence Prediction (NSP), the model learns whether two sentences are logically connected, which improves its understanding of sentence relationships and context flow.

In this research, BERT is applied to generate semantic representations of learner responses and reference answers. These representations are then compared using semantic similarity analysis to evaluate conceptual understanding more effectively than conventional keyword-matching approaches.

By emphasizing contextual meaning rather than exact word overlap, the proposed system can provide a more reliable evaluation of descriptive answers.



B. Mathematical Foundation

Input Representation

In BERT, every input token is first converted into a dense vector representation before entering the Transformer layers. The overall input representation is formed by combining three different embeddings:

X=
$Token\ Embedding + Position\ Embedding + Segment\ Embedding$

Embedding

Token embeddings capture the semantic information of words, positional embeddings encode the order of tokens within the sequence, and segment embeddings distinguish between different sentences provided in the same input.

1Self-Attention Calculation
To analyze contextual relationships among words, BERT projects the input representation into Query (Q), Key (K), and Value (V) vectors using trainable weight matrices:

$$Q = XWQ, K = XWK, V = XWV$$

The matrices Wk,WQ,WW are learned during training and help the model identify how tokens interact with one another in a sequence.

2.Attention Mechanism
The self-attention operation determines the importance of each token relative to all other tokens in the sentence:

Attention (Q, K, V) =$softmax(dkQKT)V$

First, the similarity between Query and Key vectors is computed. These scores are then normalized using the softmax function to generate attention weights. The resulting weights determine the amount of contextual information each token receives from other tokens within the sequence.

3.Multi-Head Attention

Instead of relying on a single attention computation, the Transformer architecture employs multiple attention heads simultaneously to capture different semantic and contextual relationships within the input sequence.

MultiHead(Q, K, V) =$Concat\ (head1, \ldots, headn)\ WO$

ach attention head captures distinct linguistic and semantic relationships, allowing the model to identify diverse contextual patterns within the input sequence.

4.Feed-Forward Network

After attention processing, each token representation passes through a fully connected feed-forward neural network:

FFN(x)=$max\ (0, xW1 + b1)\ W2 + b2$

This layer introduces non-linearity into the architecture and further refines token representations before they are forwarded to the next Transformer layer.

V. CHALLENGES IN TRADITIONAL METHODS

Conventional automated assessment systems used in e-learning platforms face several challenges that limit their ability to evaluate descriptive answers effectively. Many existing approaches depend heavily on keyword matching techniques,

where evaluation is based mainly on the presence of specific words or phrases. Although such methods are simple to implement, they often fail to understand the actual meaning and context of a student's response.

Another major limitation is the lack of semantic understanding. Traditional systems are unable to analyse deeper relationships between words and sentences, which can result in incorrect evaluation when students express correct ideas using alternative wording or different sentence structures. As a result, answers that are conceptually accurate may still receive lower scores due to limited contextual analysis.

Existing methods also provide little
support for concept-level assessment. They generally do not verify whether all important
concepts related to a question are properly covered in the response. In addition, these systems struggle to identify misconceptions or misleading information, reducing the reliability of the evaluation process.

Feedback generation is another significant challenge. In many platforms, the feedback provided to learners is brief and generic, offering limited guidance for improving conceptual understanding or correcting mistakes. Furthermore, descriptive answers that include detailed reasoning, explanations, or analytical discussion are often difficult to evaluate accurately using rule-based or keyword-driven techniques.

Scalability is also a concern in traditional assessment systems. Some approaches rely on manual evaluation or complex handcrafted rules, making them difficult to maintain and inefficient for large-scale educational platforms with a high number of users and submissions.

VI. METHODOLOGY

1. Model Development and Bert implementation

The proposed system applies BERT to evaluate descriptive answers by analysing their semantic meaning and contextual relevance. Rather than building a language model from the beginning, the framework uses a pre-trained BERT model that is further adapted for automated answer assessment tasks.

During the model development stage, learner responses and reference answers are first processed through text preprocessing and tokenization. The

processed text is then passed to the BERT model to generate contextual vector representations, commonly known as embeddings. These embeddings help the system understand the meaning of words based on them funding context, enabling more accurate interpretation of descriptive answers even when different sentence structures or vocabulary are used.

To assess answer quality, the framework compares embeddings of student responses with those of reference answers using cosine similarity analysis. A higher similarity value indicates stronger conceptual consistency between the two responses. Along with similarity measurement, the system performs concept extraction to identify important concepts included in the answer and evaluates concept coverage to determine whether essential topics have been addressed completely.

The framework also incorporates rule-based validation techniques to identify misconceptions, irrelevant statements, and incorrect information. In addition, a multi-criteria scoring mechanism is used to evaluate responses based on correctness, completeness, and relevance. The system further generates explainable feedback so that learners can better understand mistakes and improve their conceptual knowledge.

For implementation, the model is integrated into a Python-based backend using FastAPI, which supports efficient request handling and real-time response generation. Since the framework is optimized for CPU execution, it can operate effectively in scalable and resource-limited educational environments without requiring expensive hardware resources.

2.System Architecture

A. Presentation Tier (Client Interface)

The presentation layer is developed using React.js and is responsible for handling dynamic user interactions, managing interface states, and capturing descriptive responses submitted by students. It also supports multi-step curriculum and assessment workflows while enabling smooth communication with the backend services through asynchronous API requests. In addition, the interface displays real-time Explainable AI (XAI) feedback components without blocking user interaction, thereby improving responsiveness and usability.

B. Application Tier (Asynchronous Processing and NLP Pipeline)

The application layer is implemented using the FastAPI ASGI framework and serves as the central processing engine of the CHATUR framework. The architecture is optimized for scalable and computationally efficient CPU-based execution.

Text Preprocessing

This module performs preprocessing operations on raw textual responses, including normalization, tokenization, stop-word filtering, and noise removal to improve downstream semantic analysis.

Transformer-Based Semantic Engine

The semantic analysis component converts textual responses into dense vector representations using pre-trained Sentence-BERT (SBERT) and DistilRoBERTa transformer models. These embeddings capture contextual and semantic relationships between student answers and reference responses.

Evaluation Engine

The evaluation module executes parallel processing pipelines that combine semantic similarity computation with keyword coverage analysis. Cosine similarity is used to compare contextual embeddings, while regex-based keyword matching validates concept coverage and essential terminology within the student response.

C. Data Tier (Relational Storage and Data Integrity)

The data layer is managed using the SQLAlchemy ORM framework and is responsible for persistent storage, relational mapping, and transactional consistency across the system.

Relational Schema Structure

The database architecture follows a hierarchical educational schema that maintains relationships among topics, concepts, questions, and reference answers.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718458

Topic → Concepts → Questions → Authoritative Answers

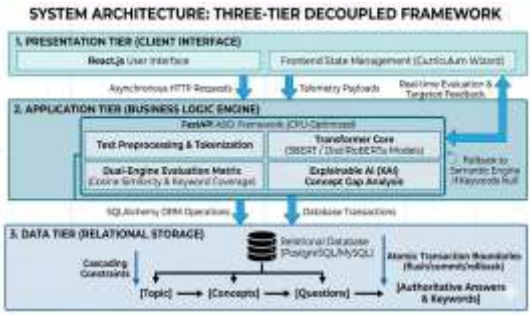
This structure enables organized curriculum mapping and efficient retrieval of educational content.

Transaction Management

To maintain database consistency during multi-step operations, intermediate updates are temporarily staged using flush operations and permanently stored only after complete validation through commit. In case of processing failure or validation errors, the system immediately performs a rollback operation to prevent partial updates or orphaned records.

Query Optimization

To improve scalability and reduce database overhead, the framework employs asynchronous eager-loading techniques such as joinedload along with aggregate query functions like func.count. These optimizations minimize excessive query generation and support efficient real-time analytics for dashboard monitoring and reporting.



3.Validation and Evaluation Metrics

The proposed CHATUR framework was evaluated using multiple semantic and performance-oriented metrics to evaluate the performance of automated descriptive answer evaluation. The framework integrates transformer-based semantic similarity analysis, keyword coverage verification, and AI-driven conceptual feedback generation to ensure accurate and meaningful assessment.

1. Semantic Similarity Score

The primary evaluation metric used in the framework is cosine similarity computed between contextual embeddings generated using transformer-based language models. The similarity between the user answer embedding (A) and the reference answer embedding (B) is calculated as:

$$\text{Similarity}(A, B) = (A \cdot B) / (\|A\| \times \|B\|)$$

Here, A denotes the embedding vector of the student's response, while B represents the embedding vector of the reference answer. The framework employs Sentence-BERT and DistilRoBERTa models to capture semantic meaning and contextual relationships between answers.

2. Keyword Coverage Validation

To evaluate conceptual completeness, the framework verifies whether important keywords and domain-specific concepts are included in the student response. This process helps identify missing concepts, partial understanding, and incomplete explanations.

$$\text{Coverage} = (\text{Matched Keywords} / \text{Total Keywords}) \times 100$$

A higher coverage score indicates that the response includes a larger proportion of the essential concepts expected in the answer.

3. Classification Performance Metrics

To measure the reliability of the automated evaluation process, the AI-generated assessment results are compared with manually verified ground-truth evaluations using standard statistical metrics.

$$\text{Accuracy} = (\text{Correct Predictions} / \text{Total Predictions}) \times 100$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1-Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Where:

- TP = True Positives
- FP = False Positives
- FN = False Negatives

These metrics collectively evaluate the accuracy and consistency of the automated grading system.

4. Concept Gap Analysis

The framework incorporates an AI-driven concept gap analysis module that identifies missing concepts, weak explanations, circular definitions, and incomplete reasoning patterns within descriptive

answers. This component combines semantic similarity analysis, concept extraction, and explainable AI-based feedback generation to provide deeper insight into student understanding.

5. Response Time Analysis
System efficiency and computational overhead are evaluated using real-time operational metrics,
including embedding generation speed, semantic evaluation runtime, and feedback generation latency. These measurements help determine the scalability and practical deployment capability of the proposed framework in large-scale educational environments.

VII.CONCLUSION

This research introduces a social learning and assessment platform integrated with an intelligent automated evaluation system for analysing descriptive answers. Conventional assessment approaches that depend mainly on keyword matching are often unable to evaluate the actual conceptual understanding demonstrated by learners. To overcome this limitation, the proposed framework utilizes BERT to perform semantic-based answer evaluation by analysing contextual meaning rather than simple word overlap.

The developed system compares student responses with reference answers using semantic similarity analysis, evaluates concept coverage, and identifies possible misconceptions within descriptive answers. In addition, the framework applies multi-criteria scoring methods and generates explainable feedback that helps learners recognize mistakes and improve their understanding of important concepts. These features contribute to a more accurate and meaningful assessment process compared to traditional automated evaluation techniques.

The platform also incorporates collaborative social learning features that allow users to create, share, and discuss educational content within an interactive environment. This combination of intelligent assessment and community-based learning helps improve user participation and encourages knowledge sharing among learners.

Furthermore, the proposed framework is designed for scalable deployment and optimized for CPU-based execution, making it suitable for educational environments with limited computational resources
. Overall, the research demonstrates how artificial intelligence and social learning can be integrated to create an efficient, transparent, and learner-focused assessment system for modern digital education platforms.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, 2019. doi: 10.18653/v1/N19-1423.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proc. EMNLP-IJCNLP, 2019. doi: 10.18653/v1/D19-1410

[3] D. Cer et al., "Universal Sentence Encoder," arXiv preprint arXiv:1803.11175, 2018.

[4] A. Vaswani et al., "Attention Is All You Need," Proc. NeurIPS, 2017. doi: 10.48550/arXiv.1706.03762.

[5] T. B. Brown et al., "Language Models are Few-Shot Learners," Proc. NeurIPS, 2020. doi: 10.48550/arXiv.2005.14165

[6] A. Joshi, R. Sharma, and P. Verma, "Automated Descriptive Answer Evaluation Using BERT and Cosine Similarity," International Journal of Creative Research Thoughts (IJCRT), vol. 11, no. 5, pp. 120–126, 2023.

[7] S. Tayal, R. Gupta, and A. Singh, "Automated Exam Paper Evaluation Using RoBERTa and Semantic Similarity Analysis," Proc. IEEE International Conference on Computing and Information Technology, 2023. doi: 10.1109/OCIT59427.2023.10431267.

[8] P. Nair and V. Uma, "Hybrid NLP-Based Assessment System Using Semantic Analysis and Explainable AI," International Journal for Multidisciplinary Research (IJFMR), vol. 6, no. 2, pp. 1–10, 2024.

[9] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[10] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proc. EMNLP System Demonstrations, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

[11] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," Proc. ACL, 2018. doi: 10.18653/v1/P18-1031.