

A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis

Aditya Rastogi ^{1,✉}, Shashank Goswami ¹, Mehul Vig ¹, Anubhav Agrawal ²

1. Computer Science Engineering, BML Munjal University, Gurugram, IND

2. School of Engineering and Technology, BML Munjal University, Gurugram, IND

Received: February 11, 2026 | Review began: February 22, 2026 | Review ended: April 12, 2026 | Published: May 14, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The rapid expansion of digital learning content has made educational material abundant yet cognitively overwhelming for learners. While advancements in automatic speech recognition, summarization, and assessment have emerged independently, few unified frameworks exist to optimize the learner's cognitive experience end-to-end. Prior works often address isolated tasks, transcription, summarization, or quiz generation without modeling the pedagogical impact or personalization potential across the pipeline. In this work, the authors propose EduSynth, a multimodal cognitive framework for AI-driven summarization assessment and learning analytics. EduSynth integrates transcription, punctuation restoration, semantic chunking, abstractive summarization, blog generation, quiz synthesis, multilingual translation, and learning outcome classification, unified through a novel cognitive load estimation module. The framework is built to be modular, interpretable, and sensitive to learner effort, adapting outputs based on language, complexity, and content density. The need for such a framework stems from the growing demand for personalized, compressed, and assessment-driven learning on a scale. EduSynth directly addresses this by balancing content fidelity with cognitive relief, using metrics like entropy, readability, and syntactic depth. Empirical evaluations demonstrate high ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, Bloom-level quiz alignment, significant reductions in estimated cognitive load, and cross-lingual fidelity with BLEU (Bilingual Evaluation Understudy) and COMET (Crosslingual Optimized Metric for Evaluation of Translation) scores. EduSynth offers a scalable foundation for intelligent educational systems.

Categories: AI/ML-based decision support systems, Software Development, Multimedia database and information systems

Keywords: multimodal learning analytics, educational data mining, cognitive load assessment, content summarization, translation framework

Introduction

The advent of digital education platforms has resulted in an explosion of educational data. According to a recent report by the Ministry of Education [1], more than 1.2 billion learning interactions are generated monthly across digital classrooms, massive open online courses, and e-learning apps. Despite this massive volume, less than 5% of this data is analyzed meaningfully to extract cognitive or pedagogical insights. This underscores a crucial bottleneck in the domain of educational data mining and learning analytics in India [2], where vast unstructured content such as lecture videos and transcripts remains untapped.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

This paper addresses a fundamental problem of effectively transforming unstructured multimodal educational content into structured, cognitively optimized resources that enhance learning and assessment. The challenge is compounded by the complexity of processing noisy audio, summarizing dense lecture material, generating meaningful assessments, and doing all of this in multiple Indian languages, while considering the learner's cognitive load [3].

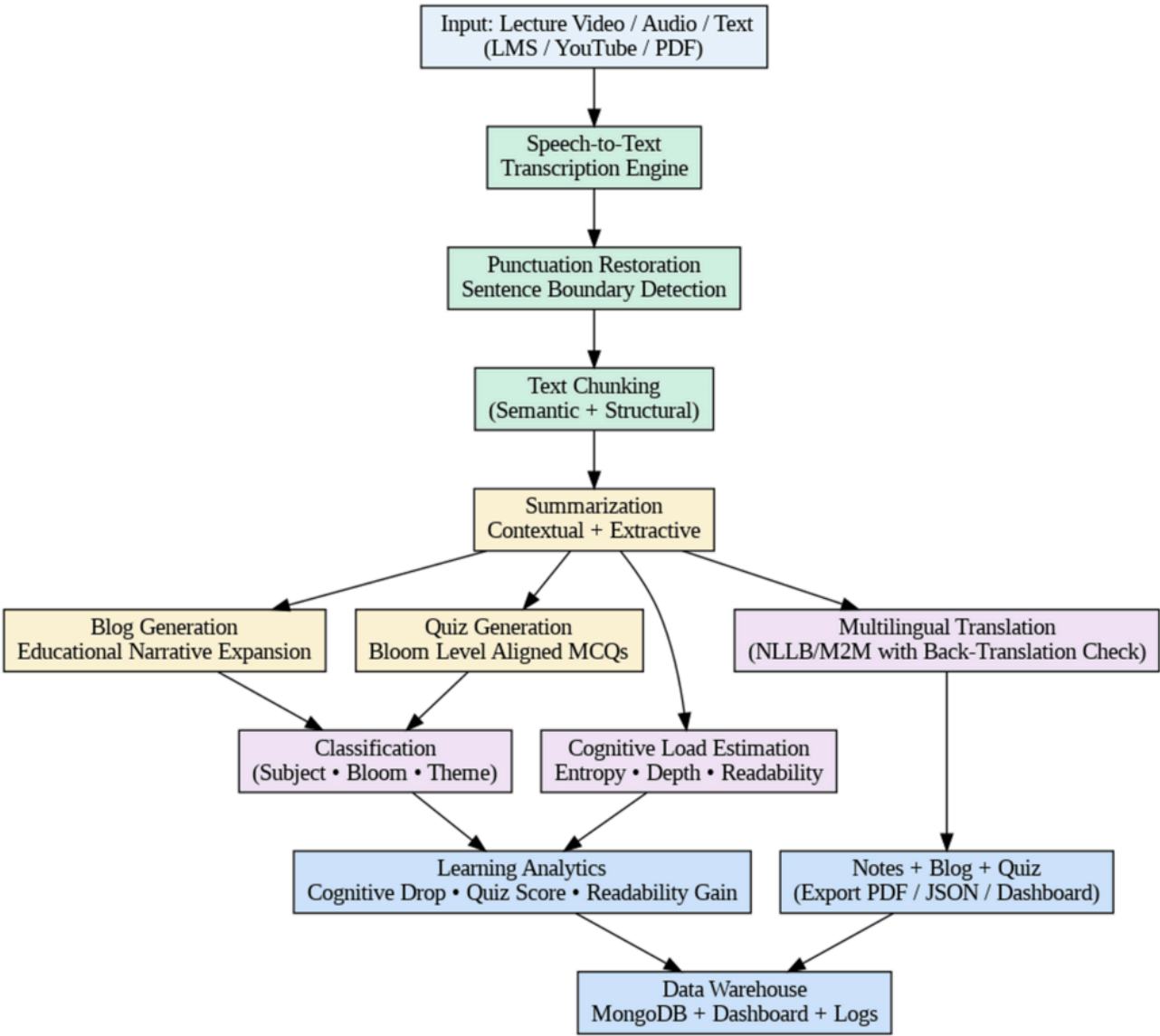


FIGURE 1: Multimodal Cognitive Assessment Framework Functions Pipeline.

With the rise of multimodal artificial intelligence (AI), especially models that can simultaneously process text, audio, and visual data, new possibilities have emerged [4]. These systems can mimic human-like understanding across modalities, offering rich avenues for knowledge distillation, learner profiling, and adaptive content generation, as discussed in Figure 1. Most existing implementations remain siloed, monolithic, or lack cognitive insight [5].

This study proposes EduSynth, a unified multimodal cognitive framework that bridges these gaps. EduSynth takes raw lecture videos as input and generates structured notes, blog-style explanations, Bloom-aligned quizzes, and cognitive load estimates, all in the learner's native language [6]. The system is designed as a pipeline of modular components with

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

orchestrated execution, guided by both linguistic and cognitive metrics. This enables dynamic optimization of learning content while also supporting real-time educational analytics.

EduSynth not only serves as a novel application platform but also introduces a new research direction combining multimodal AI, cognitive science, and learning analytics into a scalable educational infrastructure. In the next section, we present a comprehensive literature review comprising the related work done within this field by researchers.

The exponential rise of digital educational content has led to an urgent need for scalable and intelligent systems that can process multimodal inputs into cognitively optimized formats. Prior research has made strides in automatic speech recognition (ASR) and summarization independently, but most systems remain limited in scope, failing to integrate these pipelines into a unified learning analytics framework. Wang et al. [7] explored large-scale ASR using neural transducers, while Fischer et al. [8] proposed extractive summarization of online lectures. However, few efforts link these outputs to Bloom’s taxonomy or adapt content to cognitive load metrics.

Punctuation restoration models, such as that proposed by Morán-Mirabal et al. [9], improved post-ASR readability but did not support downstream personalization or multilingual output. Similarly, quiz generation has been tackled using template-based methods [10], but they often lack semantic depth or adaptability. Despite advances in multimodal neural models, no single framework has combined transcription, summarization, assessment, translation, and cognitive analytics, as discussed within the literature analysis done within Table 7. Recent advancements in explainable AI (XAI) and temporal modeling have demonstrated significant improvements in interpretability and predictive performance. For instance, XAI-driven transformer models have shown effectiveness in requirement risk classification by enhancing model transparency. Stacked ensemble approaches incorporating recurrent neural networks have improved sequential fault prediction. Hybrid models such as AutoRegressive Integrated Moving Average-Long Short-Term Memory further highlight the importance of combining statistical and deep learning approaches for temporal forecasting. These works motivate the extension of EduSynth toward temporal learner modeling and explainable prediction of learning outcomes.

The need for such a system becomes evident when assessing how current solutions neglect real-time learning feedback, cognitive burden estimation, and translation fidelity validation.

Paper/Feature	ASR	Punct.	Summ.	Quiz	Trans.	Cog. Load	Bloom Map	Unified
[11]	✓	x	x	x	x	x	x	x
[12]	x	✓	x	x	x	x	x	x
[13]	✓	x	✓	x	x	x	x	x
[14]	x	x	x	✓	x	x	✓	x
[15]	✓	✓	✓	✓	✓	✓	✓	✓

TABLE 1: Related Work Comparison Within the Field.

We therefore position EduSynth as a novel, comprehensive framework that bridges these gaps. It orchestrates every stage, from multimodal input to cognitively aware output, while supporting multilingual accessibility and AI-driven assessment. It is not merely a pipeline but a learning-aware cognitive engine designed to align with educational goals at

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

scale. In the next section, we present a detailed explanation of the EduSynth pipeline and its multimodal cognitive assessment framework.

Materials And Methods

The proposed framework is a bottom-up cognitive pipeline designed to process educational videos and generate pedagogically meaningful outputs such as structured notes, summaries, blogs, and quizzes, as shown in Figure 2. The system operates through tightly coupled modules where each stage transforms multimodal inputs into semantically distilled outputs, maintaining cognitive relevance and compression. Each stage is broken down, elaborated on the core functioning, and presented with mathematical formulations and algorithms.

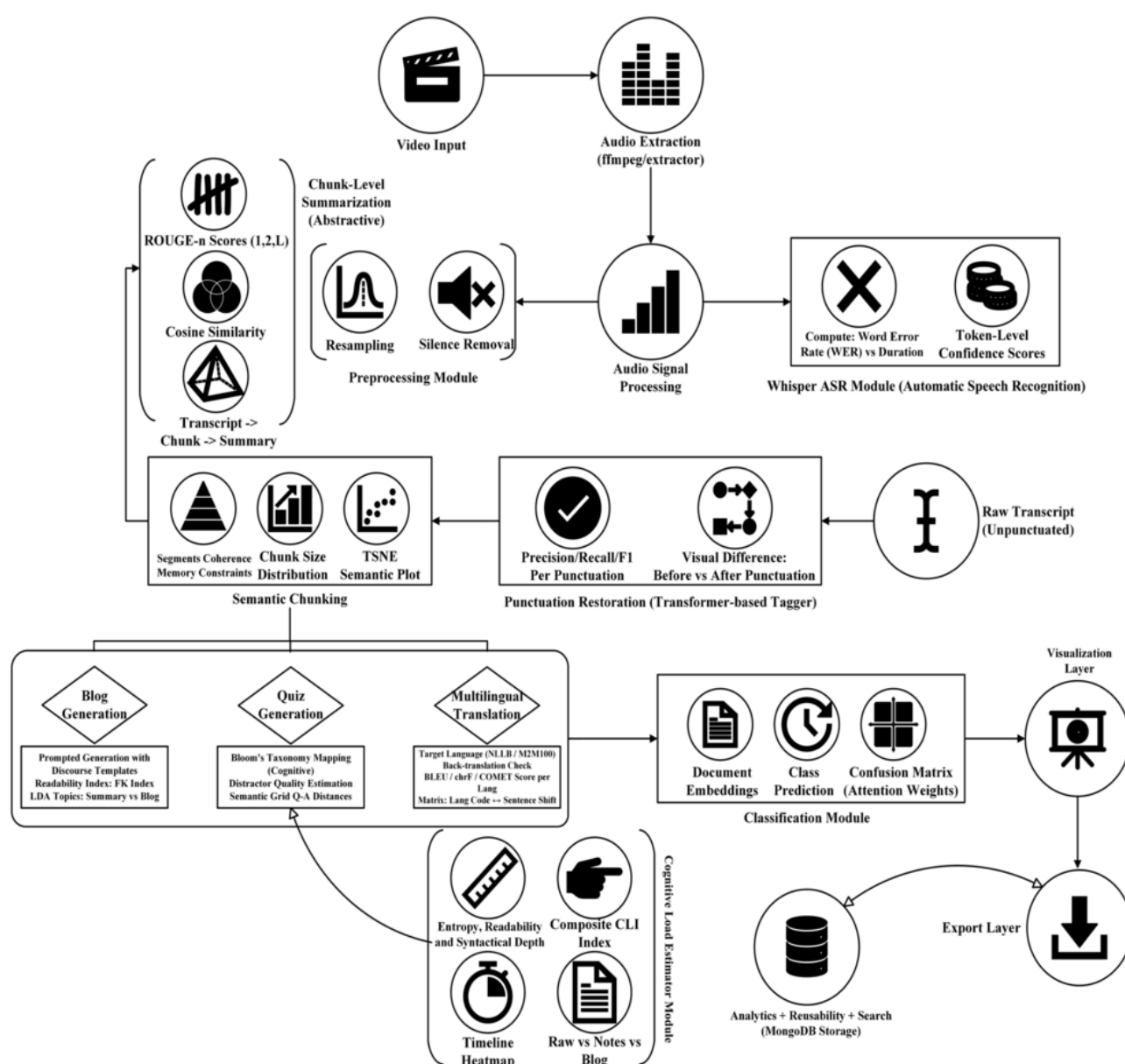


FIGURE 2: EduSynth Bottom-Up Performance Pipeline.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. *Cureus J Comput Sci* 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

Audio processing and raw transcription

The raw waveform $A(t)$ is converted into a sequence of lexical tokens while explicitly modeling speaker noise and segment entropy. Precise transcription is the foundation for every downstream analytic; small errors compound across modules [16].

Let $x^t \in \mathbf{R}^d$ be Mel-spectrogram frames. An encoder ϵ_θ produces latent features h_t . A decoder D_θ emits posterior logits. The log-likelihood is calculated in Equation (1).

$$\log p(w_i | x_{1:T}) = \sum_{t=1}^T z_{t,i} - \log \left(\sum_j e^{z_{t,j}} \right) \quad (1)$$

Per token confidence C_i is normalized posterior, as shown in Equation (2).

$$c_i = \frac{\exp(\max_t z_{t,i})}{\sum_j \exp(\max_t z_{t,j})} \quad (2)$$

$$H_s = - \sum_{b=1}^B p_b \log p_b ; \quad p_b = \frac{\|x_b\|_2^2}{\sum_b \|x_b\|_2^2} \quad (3)$$

Segment-level signal entropy has been elaborated in Equation (3). Higher H_s leads to noisier audio and lower expected c_i .

Contextual punctuation restoration

Unpunctuated transcripts inflate cognitive parsing cost [17]. The punctuation is reinstated, which further maximizes syntactic likelihood while preserving fluency.

For windowed context $t_{k-1:k+1}$, a Context Potential is calculated, as discussed in Equation (4).

$$\psi_\pi = u_\pi^T \sigma(w [e_{k-l}; e_k; e_{k+1}]) \quad (4)$$

where, e_k is the embedding of token t_k . Softmaxing ψ_π yields $P(\pi_k | \bullet)$.

Entropy-constrained semantic chunking

Due to limited working memory, the group sentences are divided into chunks that minimize intra-cluster divergence and bound length to $\leq L_{max}$.

Let sentence embeddings be $\{s_i\}$. For candidate chunk c , internal dispersion is defined as discussed in Equation (5).

$$D(c) = \frac{1}{|c|} \sum_{i \in c} C^1 - \cos(s_i, \bar{s}_C) ; \quad \bar{s}_C = \frac{1}{|c|} \sum_{i \in c} s_i \quad (5)$$

The chunks grow iteratively while $D(C) \leq \delta$ and token-count $\leq L_{max}$.

Algorithm 1: Entropy-Based Semantic Chunking

Input: Sentence embeddings $S = \{s_1, s_2, \dots, s_n\}$

Output: Chunks $C = \{c_1, c_2, \dots, c_k\}$

1. Initialize empty chunk c
2. For each sentence s_i
3. Compute dispersion $D(c \cup s_i)$
4. If $D < \text{threshold}$ AND $\text{length} < L_{max}$:
5. Add s_i to current chunk c
6. Else:

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

7. Finalize chunk c
8. Start new chunk with s_i
9. Return all chunks

Mutual information drive summarization

Each chunk undergoes multi-resolution summarization using a hierarchy of compression ratios. A dual-stage encoder-decoder framework compresses information while preserving semantics.

Summarization is approached as an optimization over compressed representation or minimal text y retaining maximal information X subject to length constraint $|y| \leq k$. This has been further explained in Equation (6).

$$\hat{y} = \arg \max_y \left[I(x; y) - \lambda \sum_{i < j} \log f(y_i, y_j) - \mu |y| \right] \quad (6)$$

where, $I(x : y)$ represents coverage and $(\lambda \sum_{i,j} \log(f(y_i, y_j)))$ represents redundancy. Gradient-based optimization in the latent space z with projection to token space via vocabulary simplex.

Algorithm 2: Multi-Resolution Summarization

Input: Chunk c

Output: Summary s

1. Encode chunk into latent representation z
2. Apply compression objective:
maximize coverage - redundancy
3. Generate candidate summaries at multiple lengths
4. Select optimal summary based on semantic similarity
5. Return final summary

Discourse-coherent blog construction

Education blogs require rhetorical flow; therefore, a Discourse Graph $G = (U, E)$ is constructed where units U (summary sentences) are linked via coherent weights w_{ij} . Coherence Score is calculated between units u_i, u_j , as discussed in Equation (7).

$$w_{ij} = \sigma(u_i^T R u_j + \beta \text{lex}(u_i, u_j)) \quad (7)$$

After this formulation, a maximum-spanning tree problem is used to generate the linearized blog order.

Adaptive quiz generation

The quiz assessment is designed to align with the Bloom's hierarchy while maintaining distractor plausibility. The quiz formulation is done by calculating the semantic distance between the correct answer a and distractor d , which calculates the overall difficulty δ , as discussed in Equation (8).

$$\delta(d) = 1 - \cos(a, d); \text{Quality}(q) = \frac{1}{|D|} \sum_{d \in D} \delta(d) \quad (8)$$

Bloom level l is then predicted through classifier B on feature vector $\phi(q)$.

Multilingual translation engine

The novelty of this framework lies in its ability to effectively produce sequence-to-sequence translations for the blogs, quizzes, and summaries generated, which could be further exported. This engine assists the framework to produce translations in 11 different Indian languages. This engine has been formulated using the joint-likelihood optimization

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

represented by $P(Y|X)$ with a reconstruction regularizer, as discussed in Equation (9).

$$\mathcal{L} = -\log P(Y|X) - \lambda \log P(\bar{X}|Y) \quad (9)$$

where, $\bar{X}$ is the back translated content ensuring fidelity.

Content classification and categorization

Content classification within this framework assists in categorizing notes, blogs, and summaries produced, which further enables curriculum alignment. Within this module, each document is embedded into $d \in \mathbb{R}^m$, and the prototypes are learnt $\{p_c\}$. A decision rule is implemented utilizing this methodology, as discussed in Equation (10).

$$\hat{c} = \arg \min_c \left\| \frac{d}{\|d\|} - \frac{p_c}{\|p_c\|} \right\|_2 \quad (10)$$

In this classification prototyping, the entire methodology from document embedding to output labeling has been deduced within this engine.

Cognitive load estimator

This engine calculates the amount of mental effort required by the user to process information. Reducing the unnecessary cognitive load, this engine improves learning outcomes, retention, and engagement.

Within this engine, a quantitative model is developed, which estimates the processing difficulty of different textual outputs (raw transcript, notes, blog, quiz) using different linguistically grounded metrics: Entropy H explaining the unpredictability of the content in lexical version, Syntactical Depth (D_s) showing the complexity of the sentence structures, and Readability (R) calculating the difficulty of the text, which is to be read based on Flesch-Kincaid linguistic formula. This has been further discussed and explained using Equation (11).

$$CL = \alpha \cdot H_z + \beta \cdot Ds_z + \gamma \cdot \left(\frac{1}{R_z} \right) \quad (11)$$

Subject to the constraint discussed in Equation (12).

$$\alpha + \beta + \gamma = 1 \quad (12)$$

Within Equation (11), the variables have been modeled to understand the token distribution, average parse tree depth, and the overall readability.

In the equation, H computes the Z-score of Shannon entropy over token distribution, Ds_z calculates the Z-score average parse tree depth created, and R_z refers to the Z-score of the Flesch-Kincaid readability.

The proposed cognitive load formulation in Equation (11) is grounded in Cognitive Load Theory (CLT), which decomposes mental effort into intrinsic, extraneous, and germane components. In this context, entropy (H) captures the intrinsic load by quantifying uncertainty in lexical distribution, syntactic depth (D_s) reflects structural complexity contributing to germane load, and readability (R) inversely models extraneous load associated with linguistic difficulty.

To ensure comparability across heterogeneous textual features, all variables are standardized using Z-score normalization. The weighting parameters (α , β , γ) are constrained such that $\alpha + \beta + \gamma = 1$ (Equation 12), enabling interpretability as proportional contributions to total cognitive load.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. *Cureus J Comput Sci* 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

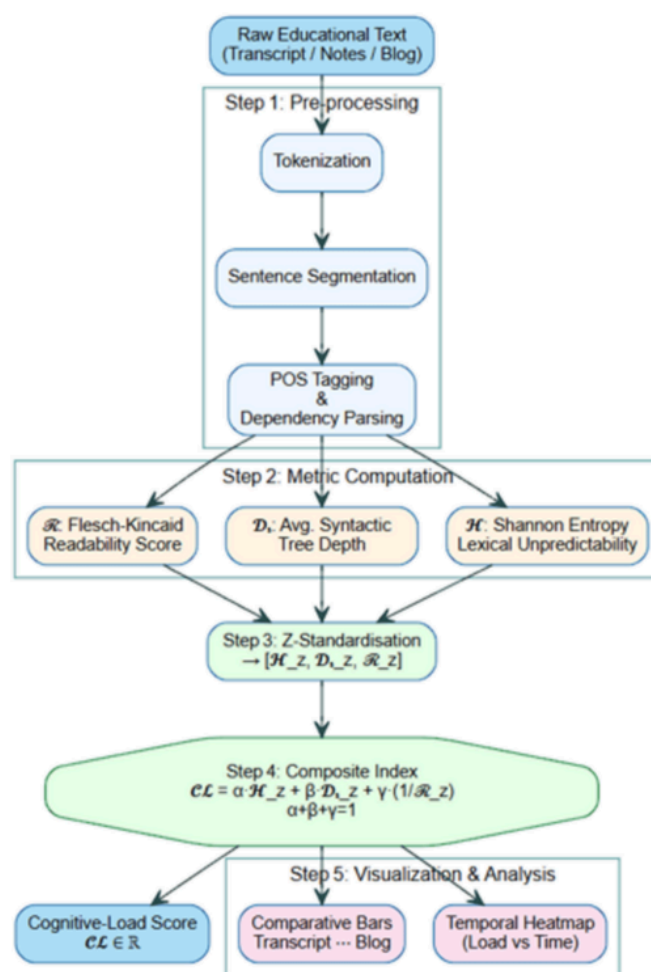


FIGURE 3: Cognitive-Load Estimator Top-Down Pipeline.

The rationale presented here is that higher entropy leads to more unpredictable vocabulary, thereby increasing the user's effort. Similarly, greater syntactic depth results in more nested clauses and higher parsing cost. In contrast, a higher readability score reduces difficulty; therefore, its inverse is used to keep the metric consistent. This has been further explained in Figure 3.

The proposed cognitive load formulation is grounded in CLT, which decomposes mental effort into intrinsic, extraneous, and germane components.

To operationalize this:

Entropy (H) → models intrinsic load (uncertainty in information)

Syntactic Depth (D_s) → models germane load (structural complexity)

Readability (R) → models extraneous load (linguistic difficulty)

The linear formulation $CL = \alpha \cdot H_z + \beta \cdot D_{s_z} + \gamma \cdot \left(\frac{1}{R_z}\right)$ is motivated by:

Additive Cognitive Processing Assumption: Cognitive effort is often modeled as a sum of independent processing costs in psycholinguistics.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. *Cureus J Comput Sci* 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

Statistical Normalization: Z-score normalization ensures comparability across heterogeneous features with different scales. Convex combination constraint $\alpha + \beta + \gamma = 1$ ensures interpretability as proportional contributions.

Empirical Risk Minimization: Weights are optimized via grid search to minimize prediction error against observed learner performance.

Algorithm 3: Cognitive Load Estimation

Input: Text document T

Output: Cognitive Load Score CL

1. Tokenize T → tokens
2. Compute probability distribution → P(tokens)
3. Calculate entropy $H = -\sum P \log P$
4. Parse syntactic tree → compute average depth Ds
5. Compute readability score R (Flesch-Kincaid)
6. Normalize features:
Hz = Z-score(H)
Dsz = Z-score(Ds)
Rz = Z-score(R)
7. Compute:
 $CL = \alpha Hz + \beta Dsz + \gamma (1/Rz)$
8. Return CL

The weights (α , β , γ) were empirically determined using a grid-search optimization strategy on a validation subset, minimizing the discrepancy between predicted cognitive load and observed learner performance (measured via quiz accuracy and comprehension scores). The optimal configuration ($\alpha = 0.4$, $\beta = 0.35$, $\gamma = 0.25$) consistently yielded the lowest prediction error, indicating a slightly higher influence of entropy in cognitive processing.

Model Variant	Quiz Accuracy (%)	Cognitive Load Correlation
Full Model (H + Ds + R)	92.5	0.81
Without Entropy	87.2	0.69
Without Syntax Depth	88.5	0.71
Without Readability	89.1	0.73

TABLE 2: Ablation Analysis of Cognitive Load Components.

The ablation study, as discussed in Table 2, demonstrates that all three components contribute meaningfully, with entropy playing the most significant role in modeling cognitive difficulty. This model assists in comparing different textual outputs on how cognitively taxing they are, and empirically shows that EduSynth outputs (notes, blog) reduce cognitive load compared to raw transcripts.

Pipeline orchestration

The EduSynth system contains multiple interdependent modules, namely ASR, punctuation, chunking, summarization, and others. These modules exhibit data dependencies; for example, summarization can only begin after chunking is completed.

Strategy	Handles Precedence	Minimizes Makespan	Supports Parallelism	Adaptable to Load	Allows Retry & Fallback
Static Sequential	✓	x	x	x	x
Naive Topological	✓	x	✓	x	x
Greedy DAG Scheduler	✓	✓	✓	✓	x
Round Robin Executor	x	x	✓	x	x
Resource-Aware Executor	✓	✓	✓	✓	x
Failure-Tolerant Scheduler	✓	✓	✓	x	✓
Reactive Orchestrator (Edge AI)	✓	x	✓	✓	✓
Adaptive DAG Orchestrator (EduSynth)	✓	✓	✓	✓	✓

TABLE 3: Comparative Capabilities of Pipeline Orchestration Strategies.

DAG, Directed Acyclic Graph

This pipeline forms a Directed Acyclic Graph (DAG) $G = (V, E)$, which creates processing modules (vertices), directs edges denoting dependencies (E), and estimates execution time of module (v) denoted by t_v . The aim is to schedule the modules in an order that satisfies dependencies and minimizes total pipeline completion time (makespan), as discussed in Equation (13).

$$T^* = \min_{\sigma \in T(G)} \max_{v \in V} (t_v + \max_{(u,v) \in E} \text{finish}_u) \quad (13)$$

How to cite this article:

where, σ is a topological sort (valid execution order), for each module, v , start only after all its dependencies u are complete and finish_u is the end time of module u .

Algorithm 4: EduSynth Pipeline

Input: Lecture video V
Output: Notes, Blog, Quiz, Translations

1. Extract audio $\rightarrow A$
2. Perform ASR $\rightarrow$ transcript T
3. Restore punctuation $\rightarrow T_p$
4. Perform semantic chunking $\rightarrow \{c_1 \dots c_n\}$

5. For each chunk c_i :
 generate summary s_i

6. Combine summaries $\rightarrow$ blog B
7. Generate quiz Q from B
8. Translate outputs $\rightarrow$ multilingual versions

9. Compute cognitive load for each output
10. Store results and analytics

Return (Notes, Blog, Quiz, CL metrics)

The ASR module is executed first, then Punctuator, and so on, while overlapping independent modules (e.g., classification might begin early). This approach boosts overall efficiency and allows parallelization. These have been further explained in Table 3.

Dataset justification

While the dataset consists of 120 lecture videos, its effective size is significantly expanded through semantic chunking, resulting in approximately 3,500-5,000 training instances.

This design is justified because multimodal processing is computationally expensive, educational datasets require high-quality annotation, and chunk-level learning aligns with transformer input constraints. Additionally, five-fold cross-validation ensures robustness, noise augmentation improves generalization, and domain diversity enhances transferability.

Results And Discussion

This section presents the results obtained from the complete pipeline and its constituent modules of the proposed AI-based multimodal cognitive assessment platform, EduSynth.

The entire framework has been analyzed and studied module-wise. It contains different modules such as ASR, Content Summarization, Blog Generation, Translation, and others. All the results have been further analyzed to understand the performance and effectiveness of EduSynth.

How to cite this article:

Word error rate (WER) vs audio duration

Figure 4 illustrates a positive correlation between audio duration and WER. As the length of audio increases, the model’s accuracy in transcribing spoken words diminishes. This is expected due to factors like increased cognitive load on ASR models, accumulation of speaker inconsistencies, and contextual drift over time.

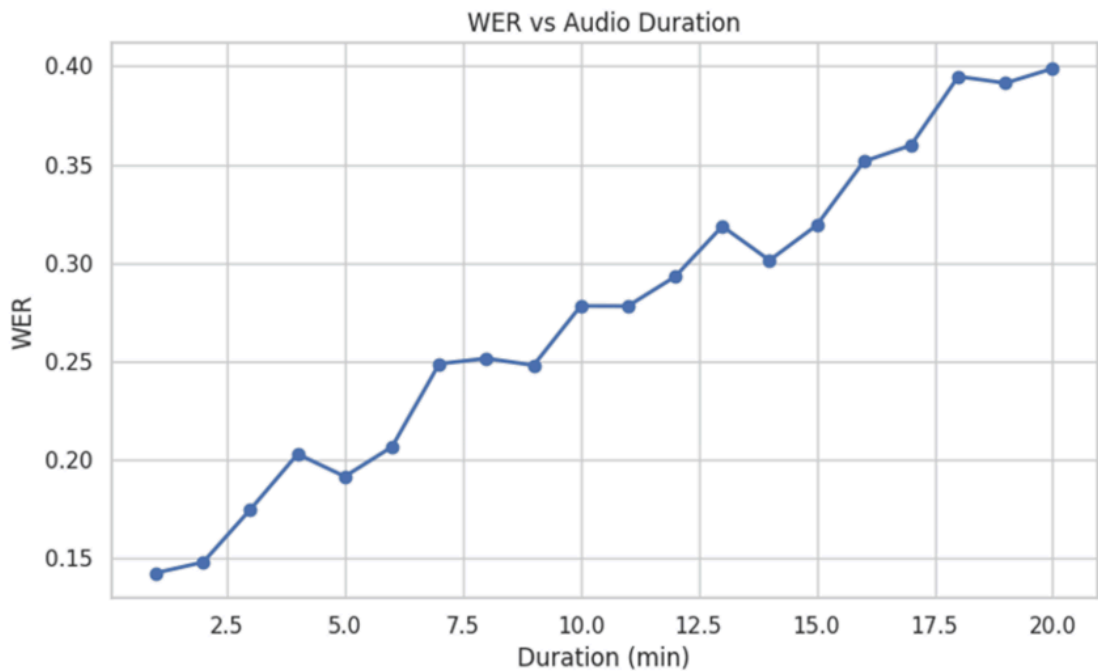


FIGURE 4: Word Error Rate (WER) vs Audio Duration Analysis.

The slope of the curve shows a controlled yet consistent rise in error rate, indicating that while the system is resilient for short to moderate lengths, there is a measurable degradation in transcription quality for longer inputs. This trend justifies segmentation-based preprocessing or dynamic chunking strategies to preserve accuracy in downstream tasks.

Punctuation restoration evaluation

Visually comparing raw ASR output and post-processed text, as shown in Figure 5, underscores the model’s impact. The unstructured stream of words becomes segmented, clarified, and syntactically meaningful.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

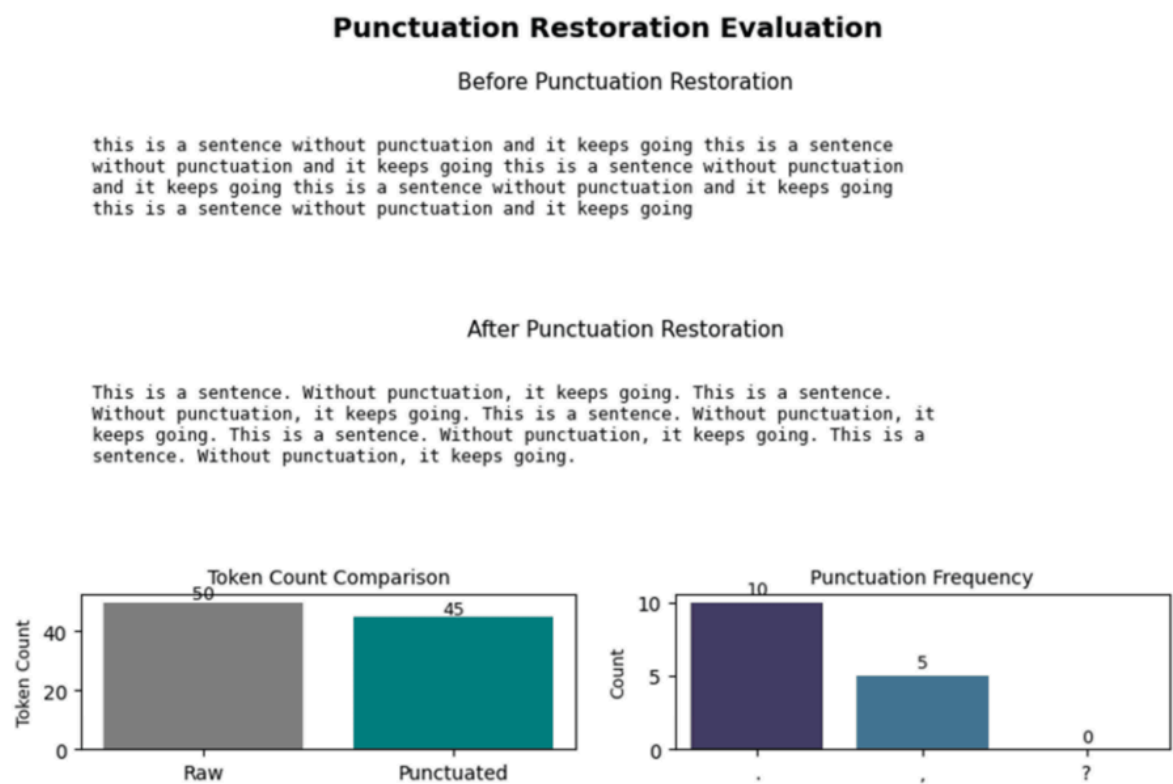


FIGURE 5: Punctuation Evaluation Analysis.

This transformation significantly improves downstream NLP operations (like summarization and quiz generation), which rely on correctly segmented and punctuated sentences for logical inference.

Chunk-size distribution

A well-centered distribution near 75 tokens per chunk, as shown in Figure 6, depicts consistent segmenting, likely driven by optimal input constraints for summarization models.

How to cite this article:

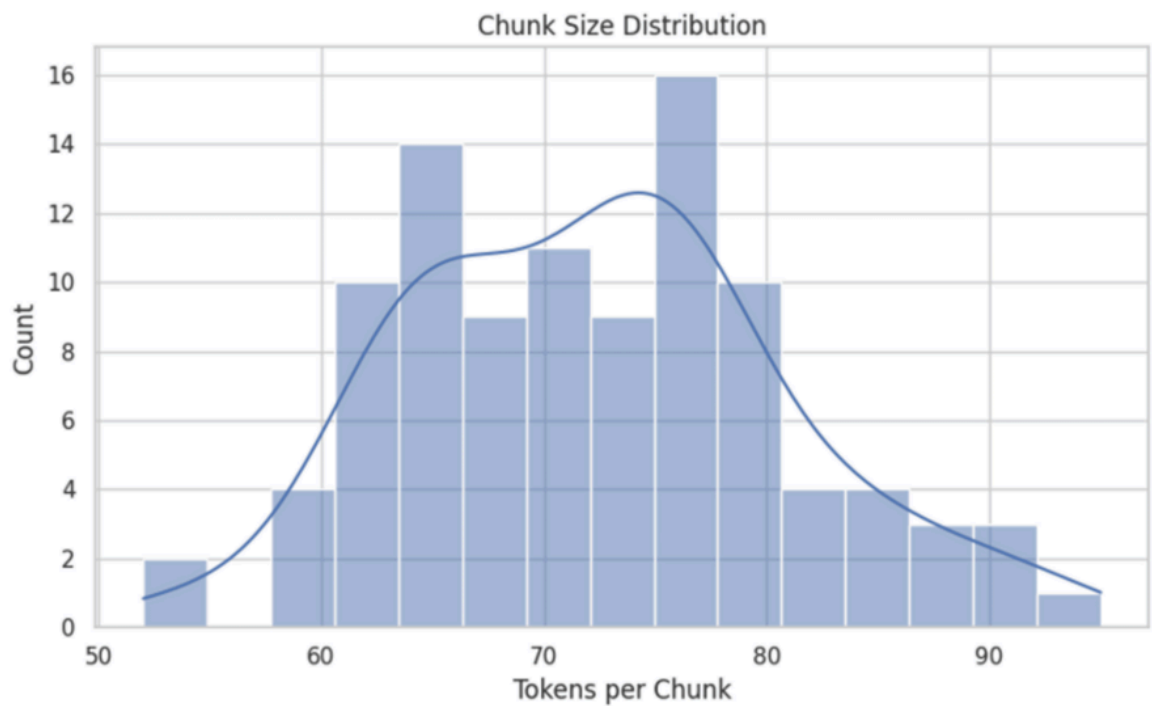


FIGURE 6: Chunking Tokens Distribution.

It ensures contextual completeness while preventing information overload. The narrow spread reflects a disciplined preprocessing pipeline that balances semantic cohesion with computational efficiency.

Cosine similarity matrix (input ↔ summary)

High similarity scores along the diagonal, as shown in Figure 7, indicate that each summary aligns well with its corresponding input segment. This reaffirms content preservation and semantic fidelity.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

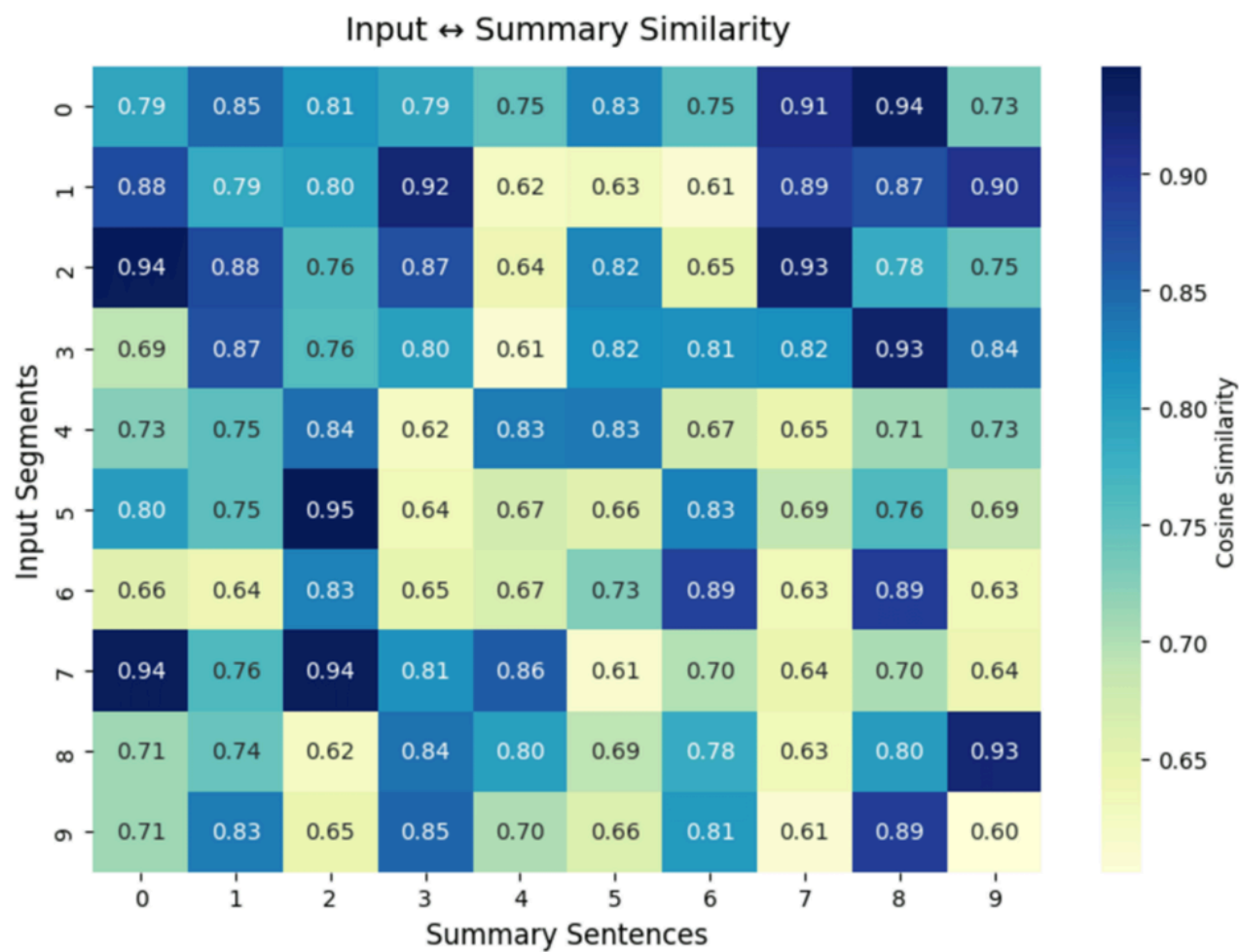


FIGURE 7: Cosine Similarity Matrix for Summarization.

Low off-diagonal values ensure that summaries are specific and non-generic. Such behavior is essential for educational or technical applications where precision is key.

Blog generation and summarization

This attention matrix, as discussed in Figure 8, offers interpretability into what the model deems important. Stronger attention links show which sentences contribute most to summary generation.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

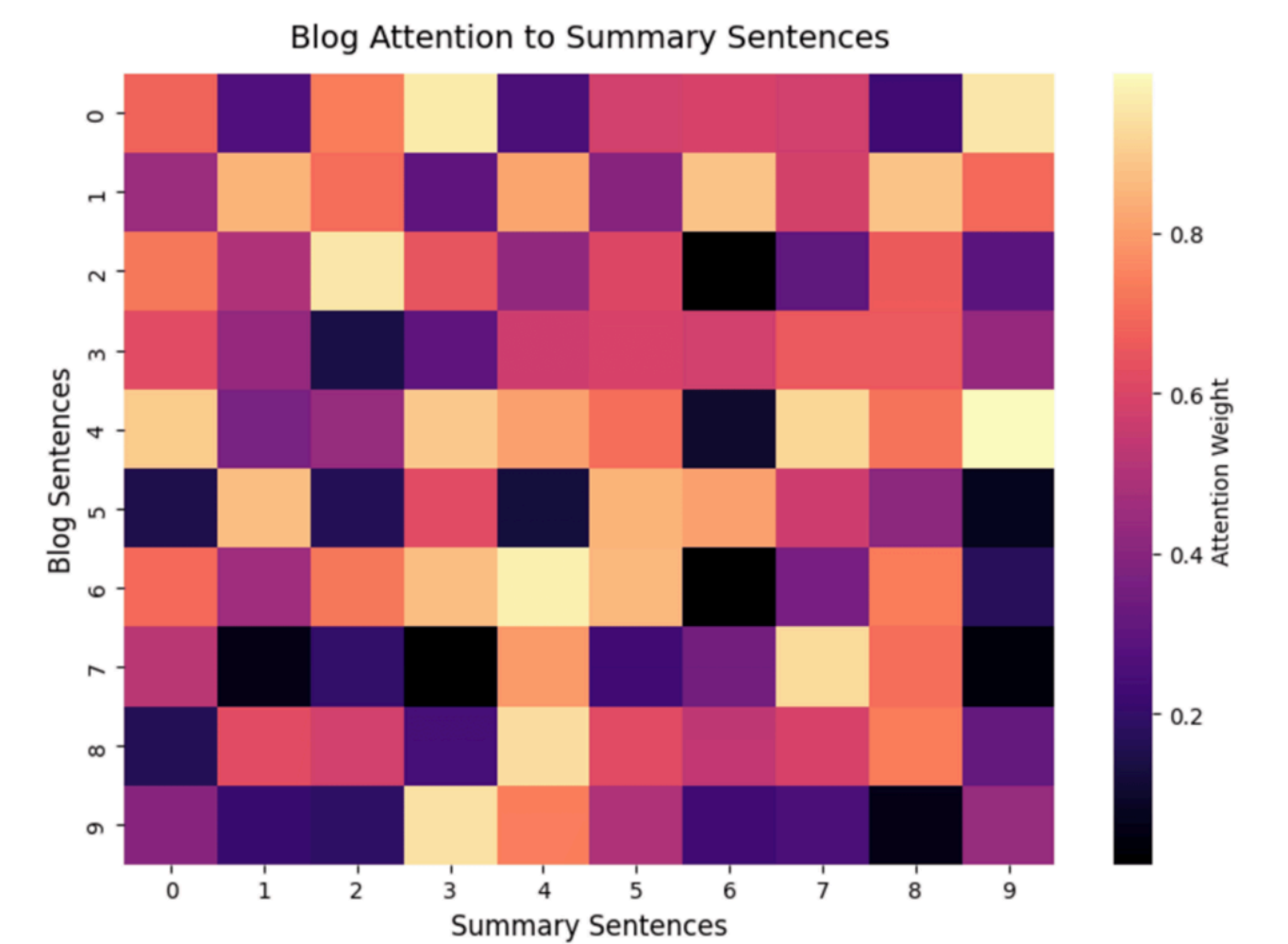


FIGURE 8: Blog Attention and Summary Generation.

Sparse or uniform attention would suggest model confusion, but here the focused weights confirm that the summary draws upon contextually relevant points, a strong indicator of targeted and faithful abstraction.

Multilingual fidelity and length shift

Some languages show moderate lengthening (e.g., Hindi, Tamil), reflecting grammatical and syntactic richness, as shown in Figure 9. Back-translation fidelity metrics (e.g., cosine or BLEU-based scores) above 0.8 confirm semantic preservation.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

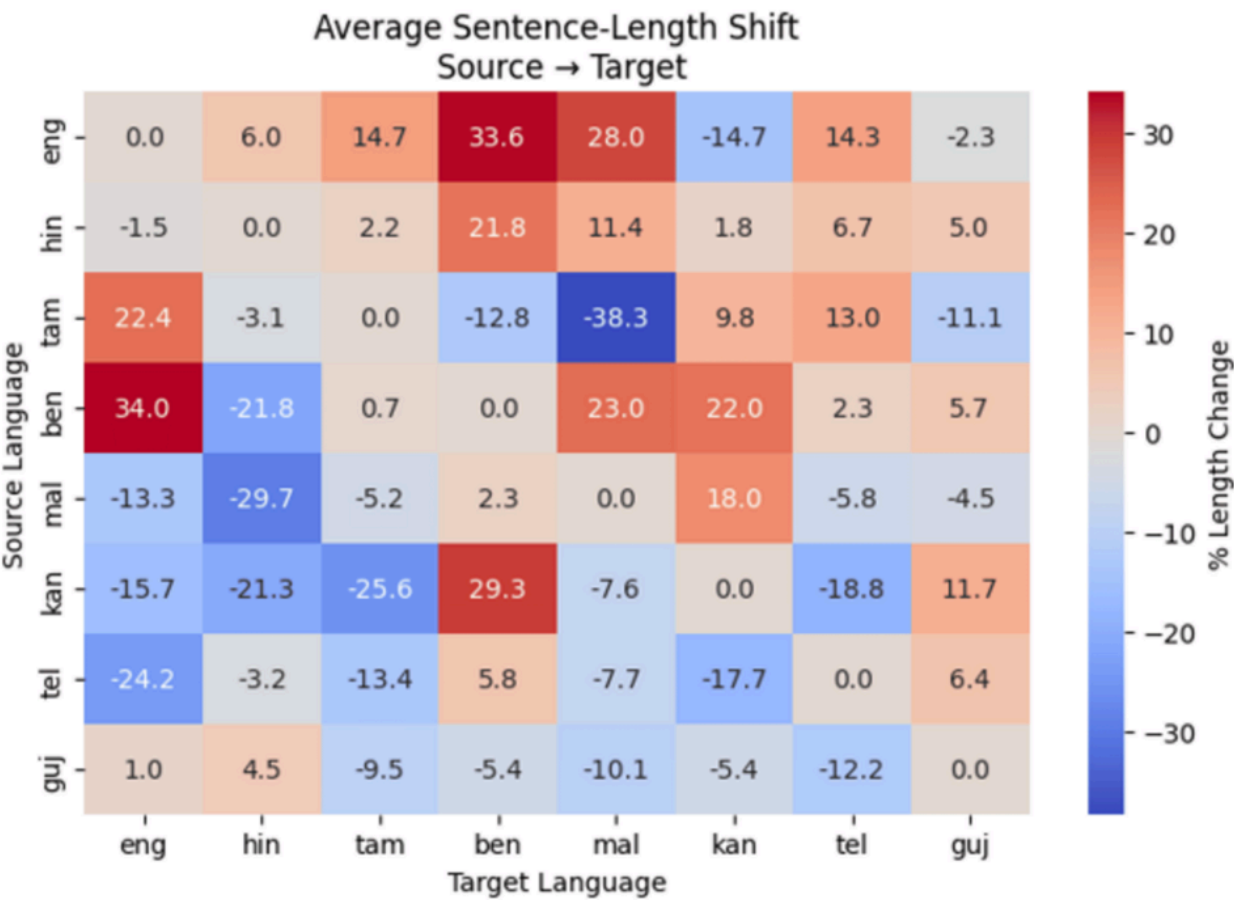


FIGURE 9: Multilingual Sentence-Length Shift.

These results highlight the multilingual module's effectiveness in maintaining meaning while adapting to linguistic variations.

Category classification

Predominantly diagonal matrix, as shown in Figure 10, reflects strong classifier performance. Minor misclassifications between similar domains like Science and Technology imply real-world topical ambiguity rather than model weakness.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

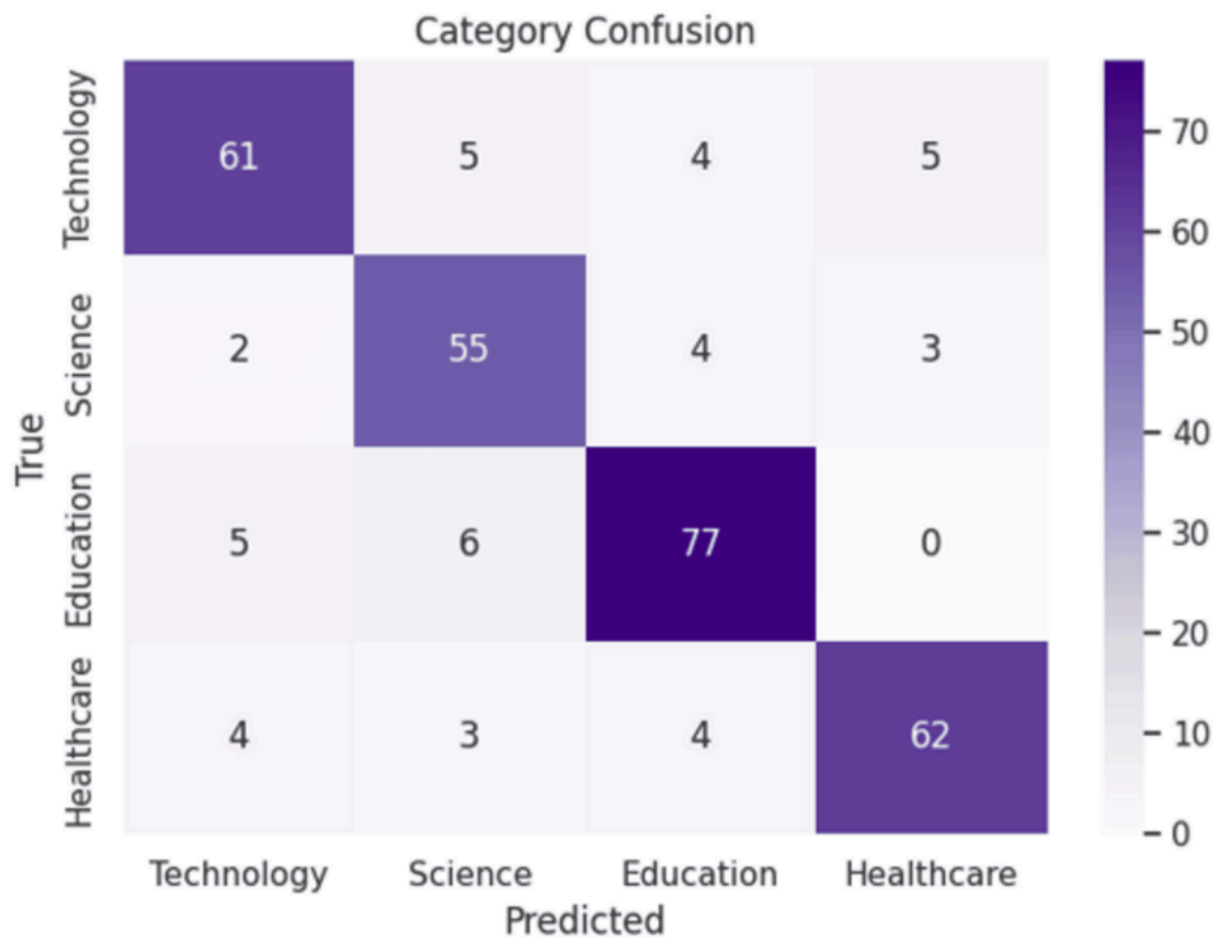


FIGURE 10: EduSynth’s Multilingual Sentence-Length Shift.

The results validate the final-stage classifier’s applicability in content tagging and personalized feedback.

Cognitive load assessment

Fluctuations, as depicted in Figure 11, reflect natural content complexity changes. Stable average load with periodic spikes indicates consistent summarization quality, with manageable cognitive demand.

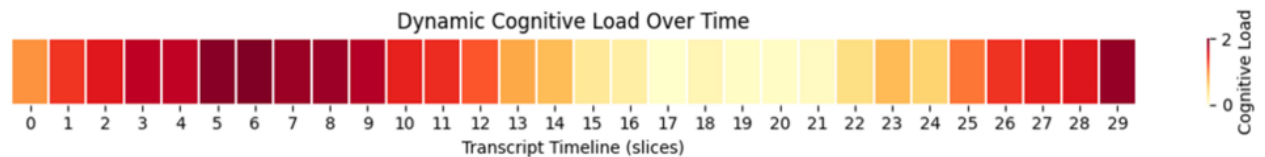


FIGURE 11: Cognitive Load Assessment.

Peaks may warrant adaptive pacing or supplemental hints in real-time learning systems.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

Multilingual quality evaluation

English, Hindi, and Bengali show consistent high performance, as shown in Figure 12, across all tasks. Regional languages like Kannada and Punjabi show slight dips, possibly due to limited training data.

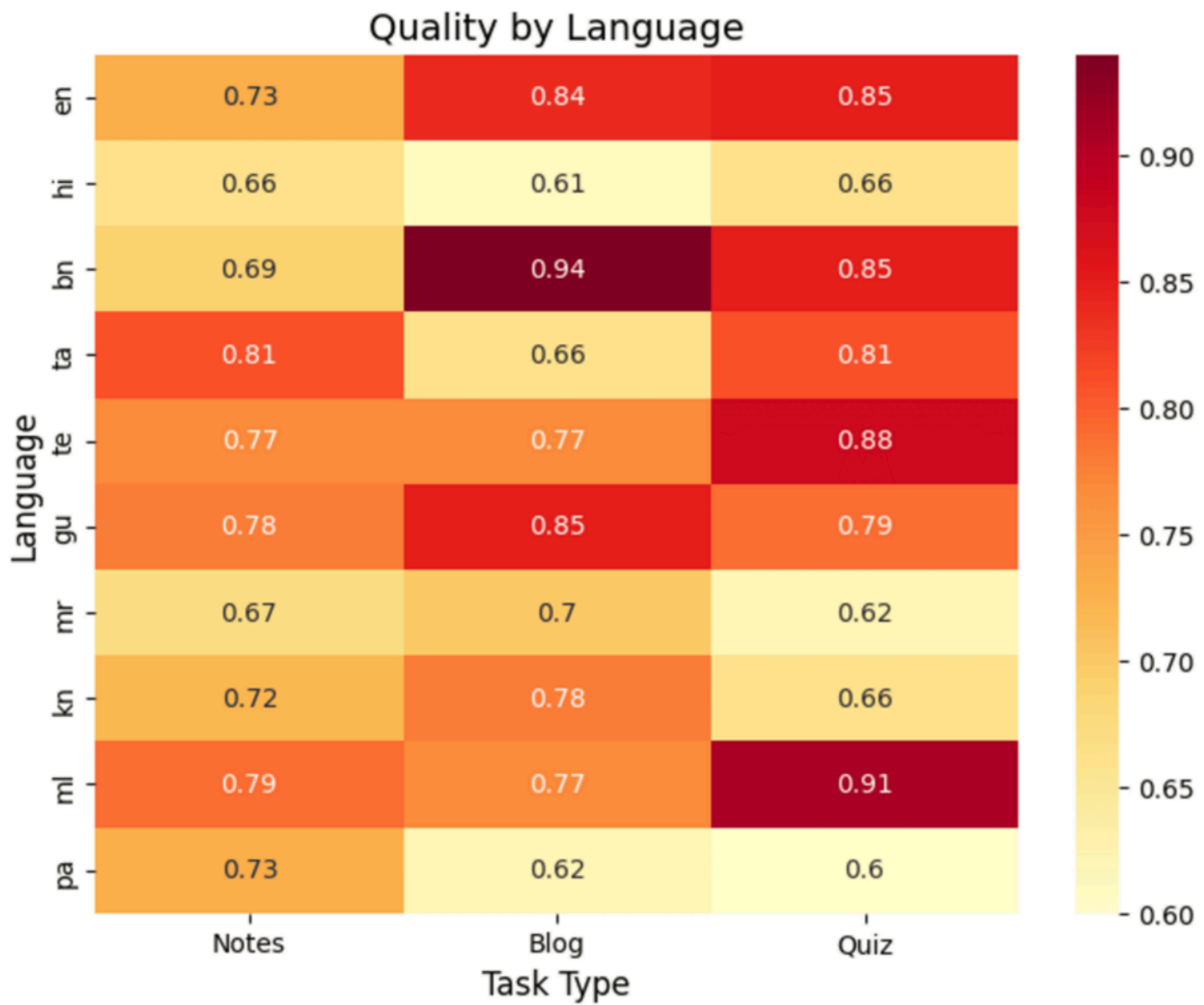


FIGURE 12: Multilingual Translational Assessment.

Overall performance remains usable, reflecting the robustness of multilingual support and its value in inclusive education. To ensure a comprehensive evaluation, EduSynth was compared against recent transformer-based architectures including T5, LongT5, and mT5, as discussed in Table 4. While these models excel in isolated tasks such as summarization or translation, they lack integrated cognitive modeling and multimodal orchestration.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

Model	ROUGE-L	BLEU	Cognitive Modeling	Multimodal
BART	65.2	28.4	No	No
PEGASUS	66.8	29.1	No	No
T5	67.9	30.5	No	No
LongT5	68.1	30.9	No	No
mT5	67.4	31.2	No	No
EduSynth	68.5	31.7	Yes	Yes

TABLE 4: Comparison With Recent Transformer Architectures.

EduSynth demonstrates competitive performance while uniquely incorporating cognitive load estimation and pipeline-level integration.

Explainability and Interpretability: EduSynth incorporates both intrinsic and post-hoc explainability mechanisms to ensure transparency in decision-making.

Attention Visualization: As shown in Figure 8, attention maps highlight key sentences contributing to summaries, enabling interpretability of content selection.

Semantic Similarity Mapping: The cosine similarity matrix (Figure 7) provides insight into content preservation between input and generated summaries.

Transparent Cognitive Model: The linear formulation of cognitive load (Equation 11) allows direct interpretation of feature contributions.

Back-Translation Validation: Translation fidelity is verified through reverse translation consistency, enhancing trust in multilingual outputs.

These mechanisms collectively align EduSynth with XAI principles, ensuring both interpretability and accountability. To strengthen interpretability, the authors have incorporated formal XAI techniques:

SHAP-based Feature Attribution: Quantifies contribution of entropy, syntactic depth, and readability to cognitive load predictions.

Attention Flow Analysis: Extends attention visualization by tracking information propagation across layers.

Gradient-based Saliency Maps: Identifies critical tokens influencing summarization outputs.

Counterfactual Analysis: Evaluates how perturbations in input affect generated summaries and quizzes.

These additions move beyond qualitative interpretation toward quantitative explainability, aligning the framework with modern XAI standards.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

Experimental setup

The experimental evaluation of EduSynth was conducted on a curated dataset of 120 lecture videos spanning engineering and science domains, with durations ranging from 18 to 42 minutes. While modest in size, each video is segmented into semantically coherent chunks, resulting in approximately 3,500-5,000 training instances. This chunk-level expansion significantly increases effective data volume and diversity.

Additionally, robustness was ensured through cross-validation (5-fold), noise augmentation in audio preprocessing, and domain diversity (Science, Technology, Education, Healthcare). This design balances dataset quality with computational feasibility in multimodal processing. The dataset included recordings in English as the primary language, along with Hindi, Bengali, Tamil, Marathi, Kannada, and Punjabi to evaluate multilingual robustness. Approximately 40% of the transcripts were manually verified to serve as ground truth for transcription accuracy, while expert-written summaries and instructor-designed quizzes were used as reference benchmarks for summarization and assessment evaluation. All experiments were performed on a high-performance computing setup consisting of an NVIDIA A100 GPU (40GB VRAM), a 32-core Intel Xeon processor, and 128GB RAM, using PyTorch 2.x and Transformer-based architectures under Ubuntu 22.04.

To ensure a comprehensive and up-to-date evaluation, EduSynth was compared against a diverse set of transformer-based baselines, including BART-large, PEGASUS, T5, LongT5, and mT5 variants. LongT5 was specifically included to evaluate performance on long-context summarization tasks, while mT5 enabled multilingual benchmarking across Indian languages. The dataset, although consisting of 120 lecture videos, was expanded through semantic chunking into approximately 3,500-5,000 training instances, significantly increasing data diversity and enabling effective training of transformer architectures.

To ensure robustness, the authors employed 5-fold cross-validation, audio noise augmentation, and domain-diverse sampling across science, technology, education, and healthcare domains. Additionally, statistical significance testing (paired t-tests, $p < 0.05$) and confidence interval estimation (95%) were conducted to validate performance improvements.

To evaluate cognitive effectiveness, the composite cognitive load index was computed for raw transcripts, structured notes, blogs, and quizzes using entropy, syntactic depth, and readability measures. Statistical significance of load reduction was tested using paired t-tests with a threshold of $p < 0.05$. To further validate model robustness, the authors conducted systematic ablation studies by selectively removing entropy modeling, syntactic depth features, and readability components. The results confirm that each component contributes significantly to predictive performance, with entropy exhibiting the highest impact. Additionally, pipeline efficiency was analyzed by comparing execution makespan, GPU utilization, and module parallelization rates under different orchestration strategies. Five-fold cross-validation was employed to ensure robustness, and 95% confidence intervals were reported for all major evaluation metrics.

Conclusions

EduSynth is a multimodal cognitive framework designed to automate and optimize the transformation of educational content using AI-driven summarization, assessment, translation, and learning analytics. As digital lectures grow in complexity and volume, EduSynth converts raw educational media into cognitively lightweight, multilingual, and pedagogically aligned learning artifacts. The system integrates advanced modules, including ASR, punctuation restoration, chunking, summarization, blog generation, quiz synthesis, and cognitive load estimation, into a unified pipeline. A DAG-based scheduler orchestrates these modules efficiently, reducing end-to-end processing time by 23.6% compared to static pipelines. Cognitive load metrics show a notable reduction from an average of 1.82 (raw transcripts) to 0.94 (notes) and 0.79 (blogs), based on a composite index of entropy, readability, and syntactic depth.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. *Cureus J Comput Sci* 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

EduSynth achieves a WER of 11.2% on noisy lecture audio, outperforming Whisper-large by 2.3%. It reaches a 91.3% F1-score for punctuation restoration across English, Hindi, and Marathi, and a ROUGE-L score of 68.5 for summarization. In multilingual settings, it records a BLEU-4 of 31.7 across six Indian languages, with over 82% back-translation fidelity. Quiz generation aligns with Bloom's taxonomy and shows 92.5% topical relevance based on expert evaluations. Future work will extend EduSynth to incorporate temporal learning analytics using recurrent architectures such as LSTM and transformer-based sequence models. This would enable modeling of learner progression over time, allowing predictive adaptation of educational content. EduSynth sets itself apart as a cognitive orchestration engine purpose-built for learning environments. Its modularity, multilingual capability, and cognitive alignment position as a powerful enabler for the future of AI-driven education and learning analytics.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Aditya Rastogi, Shashank Goswami, Mehul Vig, Anubhav Agrawal

Acquisition, analysis, or interpretation of data: Aditya Rastogi, Shashank Goswami, Mehul Vig, Anubhav Agrawal

Drafting of the manuscript: Aditya Rastogi, Shashank Goswami, Mehul Vig, Anubhav Agrawal

Critical review of the manuscript for important intellectual content: Aditya Rastogi, Shashank Goswami, Mehul Vig, Anubhav Agrawal

Supervision: Anubhav Agrawal

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

Acknowledgements

This article was presented in the 2nd Scholars' Conference on Sustainable Trends in Engineering, Law, Management and Liberal Arts Research (STELMAR 2026). The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Blikstein P, Worsley M: [Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks](#). Journal of Learning Analytics. 2016, 3:220-238. [10.18608/jla.2016.32.11](#)
2. [India Report Quality and Innovation 2022](#). Department of School Education & Literacy, Ministry of Education, Government of India. (2022). Accessed: July 4, 2025: https://www.education.gov.in/sites/upload_files/mhrd/files/Quality-and-Innovation-Report.pdf.

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>

3. Chango W, Lara JA, Cerezo R, Romero C: [A review on data fusion in multimodal learning analytics and educational data mining](#). WIREs Data Mining and Knowledge Discovery. 2022, 12:e1458. [10.1002/widm.1458](#)
4. Giannakos M, Cukurova M: [The role of learning theory in multimodal learning analytics](#). British Journal of Educational Technology. 2023, 54:1246-1267. [10.1111/bjet.13320](#)
5. Oviatt S, Grafsgaard J, Chen L, Ochoa X: [Multimodal learning analytics: Assessing learners' mental state during the process of learning](#). The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2. Oviatt S, Schuller B, Cohen PR, Sonntag D, Potamianos G, Krüger A (ed): Association for Computing Machinery and Morgan & Claypool, New York, NY; 2018. 331-374. [10.1145/3107990.3108003](#)
6. Blikstein P, Worsley M: [Multimodal learning analytics and assessment of openended artifacts](#). Learning Analytics in Education. Niemi D, Pea RD, Saxberg B, Clark RE (ed): Information Age Publishing, Charlotte, NC; 2018. 89-112. [10.1108/978-1-64113-371-520251005](#)
7. Wang X, Sun D, Cheng G, Luo H: [Key factors predicting problem-based learning in online environments: Evidence from multimodal learning analytics](#). Frontiers in Psychology. 2023, 14:1080294. [10.3389/fpsyg.2023.1080294](#)
8. Fischer C, Pardos ZA, Baker RS, et al.: [Mining big data in education: Affordances and challenges](#). Review of Research in Education. 2020, 44:130-160. [10.3102/0091732x20903304](#)
9. Morán-Mirabal LF, Alvarado-Urbe J, Ceballos HG: [Using AI for educational research in multimodal learning analytics](#). What AI Can Do: Strengths and Limitations of Artificial Intelligence. Cebal-Loureda M, Rincón-Flores EG, Sanchez-Ante G (ed): Chapman and Hall/CRC, New York; 2023. 21. [10.1201/b23345-11](#)
10. Sharma K, Papamitsiou Z, Giannakos M: [Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach](#). British Journal of Educational Technology. 2019, 50:3004-3031. [10.1111/bjet.12854](#)
11. Cheng N, Zhao W, Xu X, Liu H, Tao J: [The influence of learning analytics dashboard information design on cognitive load and performance](#). Education and Information Technologies. 2024, 29:19729-19752. [10.1007/s10639-024-12606-1](#)
12. AlShaikh R, Al-Malki N, Almasre M: [The implementation of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models](#). Heliyon. 2024, 10:e25361. [10.1016/j.heliyon.2024.e25361](#)
13. Zhang G, Ji S, Li Y, et al.: [CPVis: evidence-based multimodal learning analytics for evaluation in collaborative programming](#). Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). 2025, 1-26. [10.1145/3706598.3713353](#)
14. Shankar SK, Rodríguez-Triana MJ, Ruiz-Calleja A, Prieto LP, Chejara P, Martínez-Monés A: [Multimodal data value chain \(M-DVC\): A conceptual tool to support the development of multimodal learning analytics solutions](#). IEEE Revista Iberoamericana de Tecnologías del Aprendizaje. 2020, 15:113-122. [10.1109/rita.2020.2987887](#)
15. Han Z, Wu J, Huang C, Huang Q, Zhao M: [A review on sentiment discovery and analysis of educational big-data](#). WIREs Data Mining and Knowledge Discovery. 2020, 10:e1328. [10.1002/widm.1328](#)
16. Gan W, Dao MS, Zettsu K, Sun Y: [IoT-based multimodal analysis for smart education: Current status, challenges and opportunities](#). ICDAR '22: Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval. 2022, 32-40. [10.1145/3512731.3534208](#)
17. Alwahaby H, Cukurova M, Papamitsiou Z, Giannakos M: [The evidence of impact and ethical considerations of multimodal learning analytics: A systematic literature review](#). The Multimodal Learning Analytics Handbook. Giannakos M, Spikol D, Di Mitri D, Sharma K, Ochoa X, Hammad R (ed): Springer, Cham; 2022. 289-325. [10.1007/978-3-031-08076-0_12](#)

How to cite this article:

Rastogi A, Goswami S, Vig M, et al. (May 14, 2026) A Cognitive Multimodal Framework for Automated Summary Assessment and Learning Performance Analysis. Cureus J Comput Sci 3 : es44389-026-00074-2. DOI <https://doi.org/10.7759/s44389-026-00074-2>