# International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

*Research Paper*

# EduAssist: Evaluating a Locally Deployed Large Language Model for Educational Document Summarization

Mohammed Afzal[1], Shifa Tahreem[2]

[1]Assistant Professor, Department of CSE (Artificial Intelligence & Machine Learning), Sphoorthy Engineering College, Nadergul, India

[2]M.Tech Student, Department of CSE (Artificial Intelligence & Machine Learning), Sphoorthy Engineering College, Nadergul, India

**Abstract** —The increasing use of large language models (LLMs) for educational text processing has created opportunities for automatic summarization of lengthy learning materials. However, many LLM-based applications rely on cloud-hosted services, while the performance and computational behavior of locally deployed language models for educational document summarization remain comparatively underexplored. This study presents EduAssist, an experimental framework for evaluating a locally deployed Qwen3:1.7B model for educational document summarization. The model was executed through the Ollama runtime using a chunk-based summarization and consolidation pipeline. Experiments were conducted on 12 English-language educational samples covering topics in data mining, machine learning, artificial intelligence, and knowledge-based systems. Evaluation combined text-reduction and efficiency measures with lexical, semantic, and model-assisted assessment using compression ratio, estimated reading-time reduction, processing time, source-reference ROUGE, BERTScore, and LLM-as-a-Judge. The generated summaries achieved a mean compression ratio of 50.83%, reducing the mean source length from 1,074.17 to 478.42 words and yielding an estimated mean reading-time saving of 2.98 min. Mean ROUGE-1, ROUGE-2, and ROUGE-L scores were 0.4917, 0.2371, and 0.2892, respectively, while the mean BERTScore F1 was 0.8264 and the mean LLM-as-a-Judge score was 7.96/10. Summary generation required an average of 194.33 s per sample. Exploratory analysis further showed that greater compression was associated with lower source-reference lexical retention, whereas source-summary BERTScore F1 values remained comparatively stable across the evaluated samples. Overall, the findings provide exploratory empirical evidence for the feasibility of local educational document summarization using the evaluated Qwen3:1.7B configuration, while highlighting the need for larger datasets, model comparisons, independent reference summaries, and human evaluation.

**Keywords -** EduAssist, educational document summarization, large language models, local LLM, Qwen3, Ollama, ROUGE, BERTScore, LLM-as-a-Judge.

## 1. INTRODUCTION

The increasing use of digital learning resources has made large volumes of educational content available to students, often through lengthy PDF documents. Automatic text summarization, a Natural Language Processing (NLP) task, addresses this information load by producing concise representations of source documents while retaining important information. Recent advances in Large Language Models (LLMs) have further enabled context-aware abstractive summarization of educational content.

Existing research has explored transformer- and LLM-based summarization in educational settings [1]–[4]. Summary evaluation commonly uses lexical-overlap measures such as ROUGE, while BERTScore provides a complementary contextual semantic-similarity measure [7], [8]. More recently, LLM-based evaluation has been explored as an additional approach for assessing characteristics that conventional similarity metrics may not fully capture [12]. Local LLM deployment also allows document processing to occur within the user's computing environment without relying on an external cloud-based model service [5], [6]. However, these evaluation dimensions are often examined separately, particularly in studies of locally deployed educational summarization systems.

Educational summarization involves several considerations, including text reduction, lexical and semantic similarity, reading-time savings, and computational performance. Examining these dimensions together can provide a broader view of summarization behavior, particularly for local deployment, where compression and inference latency are important experimental and practical considerations.

This study presents EduAssist as an experimental framework for evaluating educational document summarization using Qwen3:1.7B deployed locally through Ollama. Twelve English-language educational samples are evaluated using compression ratio, estimated reading-time reduction, processing time, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and LLM-as-a-Judge. Relationships among selected input, efficiency, and evaluation measures are also examined. The study is limited to EduAssist's summarization component and a single local model configuration; therefore, it does not establish superiority over alternative local or cloud-based summarization systems.

## 2. RELATED WORK

### 2.1 Educational and LLM-Based Summarization

Recent studies have explored transformer-based and large language models (LLMs) for summarizing educational content. Zhong and Litman investigated LLM-based aspect summarization of student reflections using automatic and human evaluation [1], while Xie et al. developed an LLM-supported lecture summarization system and reported improvements in knowledge recall and student satisfaction [2]. Educational summarization has also been extended to e-book and lecture-slide content through LECTOR [3] and to Arabic science textbooks using transformer models such as AraBART, mT5, AraT5, and mBART50 [4]. Collectively, these studies demonstrate the applicability of modern language models to different forms of educational content, although their document types, languages, deployment settings, and evaluation protocols differ from the local educational-document summarization setting examined in this study.

### 2.2 Local Language Models in Education

Locally deployable language models are increasingly being investigated for educational applications because they allow processing within locally controlled computing environments. Yu et al. integrated a locally stored small language model with retrieval-augmented generation in a computing course involving more than 250 students, providing evidence for the feasibility of local language-model deployment in a course setting [5]. Jaldi et al. further investigated small and private language models for educational assessment design and highlighted both their potential and limitations in model-based evaluation [6].

These studies establish that local deployment itself is not a novel contribution. The present work instead examines a locally deployed model specifically for educational document summarization, with emphasis on summary reduction, evaluation measures, and local inference performance.

### 2.3 Evaluation of Generated Summaries

Generated-summary quality is commonly evaluated across multiple dimensions. ROUGE is widely used to measure lexical overlap, but lexical similarity alone may not capture semantic equivalence. SummEval demonstrated limitations in relying on individual automatic metrics by comparing multiple summarization measures with human judgments [7]. BERTScore complements lexical evaluation by using contextual embeddings to measure semantic similarity between candidate and reference texts [8], while QuestEval explored question-answering-based evaluation of generated summaries [9].

Factual consistency is another important consideration. Tam et al. evaluated the ability of LLMs to distinguish factually consistent and inconsistent summaries [10], while Zhong and Litman examined factual-consistency evaluation for long-document summarization and highlighted continuing challenges in factuality assessment [11]. These findings indicate that high lexical or semantic similarity should not by itself be interpreted as evidence of factual correctness.

LLMs have also increasingly been used as evaluators. Li et al. reviewed the **LLM-as-a-Judge** paradigm and identified its use in scoring and ranking generated outputs, while emphasizing limitations related to bias, robustness, and reliability [12]. Accordingly, LLM-based judging in this study is treated as a complementary evaluation signal rather than an independent ground truth.

### 2.4 Research Gap

The literature demonstrates progress in educational summarization, local language-model deployment, and multidimensional summary evaluation. However, these areas are often examined under different experimental settings. Educational

summarization studies commonly focus on specific learning resources or outcomes, while local-model studies frequently address retrieval, question answering, or assessment generation. Much of the summarization-evaluation literature reviewed here also focuses on general-purpose datasets rather than locally generated educational summaries.

This study examines their intersection by evaluating a locally deployed Qwen3:1.7B configuration for educational document summarization using compression, estimated reading-time reduction, source-reference ROUGE, source-summary BERTScore, LLM-as-a-Judge, and processing-time measurements, together with exploratory correlation analysis among selected measures. The contribution is therefore an integrated empirical evaluation rather than the introduction of a new language model or evaluation metric.

## 3. MATERIALS AND METHODS

### 3.1 Experimental Design and Dataset

This study adopted an exploratory experimental design to evaluate locally deployed language-model summarization of educational content. The experimental corpus consisted of 12 English-language educational samples manually selected from publicly accessible lecture and educational-note PDFs. The samples covered topics in data mining, machine learning, artificial intelligence, and knowledge-based systems and ranged from 593 to 1,847 words, with a mean length of approximately 1,074 words. Topic-level samples were treated as individual experimental units and processed independently using the same summarization pipeline. No manual editing was applied to the generated summaries before evaluation.

### 3.2 System Architecture

Figure 1 presents the experimental workflow. Educational PDF content was subjected to text extraction, cleaning, and language detection before being divided into approximately 1,000-word chunks. Each chunk was summarized locally using Qwen3:1.7B through Ollama, and the resulting chunk-level summaries were consolidated using a separate merge prompt to obtain one final summary per sample.

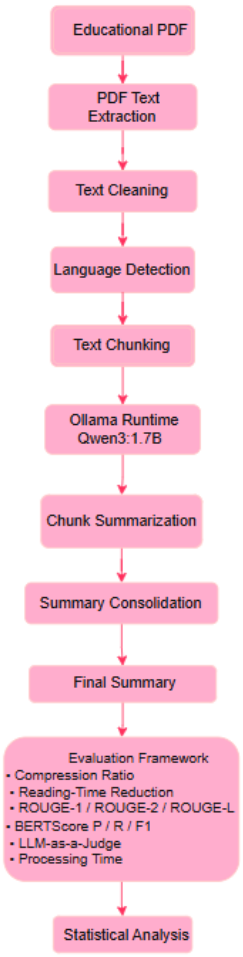


**Fig. 1**. Experimental architecture for locally deployed LLM-based educational document summarization and evaluation.

### 3.3 Local LLM Configuration and Execution Environment

All summaries were generated locally using Qwen3:1.7B through Ollama 0.32.15. The installed model reported approximately 2.0 billion parameters, a reported model context capability of 40,960 tokens, Q4_K_M quantization, and an approximate disk size of 1.4 GB. The application configured a context window of 4,096 tokens, maximum generation length of 1,000 tokens, temperature of 0, and top-p of 0.9, with thinking mode disabled.

Experiments were conducted on a 64-bit Windows system equipped with a 12th Generation Intel Core i5-1235U processor, 8 GB RAM, and an NVIDIA GeForce MX550 GPU with 2 GB dedicated graphics memory, together with integrated Intel UHD Graphics. Document processing and model

inference were performed locally without an external language-model API.

### 3.4 Summarization Procedure

Extracted text was cleaned, subjected to language detection, and divided sequentially into non-overlapping chunks of approximately 1,000 whitespace-separated words. All samples included in the experiment were identified as English. Samples exceeding the chunk size therefore produced multiple independently processed segments.

A fixed summarization prompt instructed the model to retain the main topic, important concepts, definitions, formulas, numerical values, dates, names, and technical terminology while removing repetition and avoiding information not present in the source. For multi-chunk samples, a separate consolidation prompt combined the chunk-level summaries, removed duplicate information, merged related ideas, and preserved important technical content. The final output therefore represented chunk-level summarization followed by consolidation, rather than unrestricted generation over the complete source document.

### 3.5 Evaluation Framework

Evaluation combined text reduction, estimated reading-time reduction, lexical similarity, semantic similarity, and model-assisted assessment.

For original word count $W_O$ and summary word count $W_S$, the compression ratio was calculated using Equation (1).

$$CR = \left(1 - \frac{W_S}{W_O}\right) \times 100 \qquad (1)$$

Reading time $T$ was estimated using a conservative assumed reading rate of 200 words per minute (wpm) [14], as shown in Equation (2)

$$T = \frac{W}{200} \qquad (2)$$

Estimated reading-time saving was calculated as:

$$T_{saved} = \frac{W_O - W_S}{200} \qquad (3)$$

These values represent word-count-based estimates rather than directly observed student reading behavior.

#### 3.5.1 ROUGE

Lexical overlap was assessed using ROUGE-1, ROUGE-2, and ROUGE-L F1 scores [13], with stemming enabled. The original educational sample served as the reference and the generated summary as the candidate. Consequently, the resulting scores are interpreted as source-reference lexical retention rather than conventional gold-reference summarization performance.

#### 3.5.2 BERTScore

Semantic similarity was assessed using BERTScore Precision, Recall, and F1 [8]. The implementation used the bert_score package with English specified as the evaluation language (lang="en"); no model type was explicitly specified in the application code. The original sample was supplied as the reference and the generated summary as the candidate. BERTScore was therefore interpreted as a contextual source-summary similarity measure rather than an independent measure of overall summary quality or factual correctness.

#### 3.5.3 LLM-as-a-Judge

A complementary LLM-as-a-Judge evaluation [12] instructed the model to consider five dimensions: relevance, coverage, clarity, coherence, and factual accuracy, followed by a single holistic overall score on a 1–10 scale. The overall score returned by the model was recorded; no arithmetic aggregation of separate criterion-level scores was performed.

The same locally deployed Qwen3:1.7B configuration was used for both summary generation and judging. The resulting judge scores were therefore treated as supplementary model-assisted indicators rather than independent ground truth, since same-model evaluation may introduce self-evaluation or model-preference bias.

### 3.6 Runtime and Statistical Analysis

Processing times were recorded for text extraction, language detection, cleaning, and summary generation. Summary-generation time included chunk-level generation and consolidation where required. The reported values correspond to the recorded experimental execution for each sample rather than controlled repeated benchmark trials and should therefore be interpreted as observed runtimes under the tested hardware and model configuration.

Aggregate descriptive statistics were calculated across the 12 experimental samples. Pearson correlation was used to examine linear relationships among selected variables, while Spearman rank correlation provided a rank-based comparison. Because of the small sample size, the correlations were interpreted as exploratory descriptive relationships within the evaluated dataset, and no

population-level inferential or statistical-significance claims were made

## 4. RESULTS AND DISCUSSION

### 4.1 Overall Experimental Results

The local summarization pipeline was evaluated on 12 English-language educational samples. Source lengths ranged from 593 to 1,847 words, while generated summaries ranged from 291 to 595 words, with compression ratios ranging from 10.96% to 84.24%.

Across the dataset, mean source and summary lengths were 1,074.17 and 478.42 words, respectively, corresponding to 50.83% mean compression and an estimated reading-time saving of 2.98 min. Mean BERTScore F1 was 0.8264, while average summary-generation time was 194.33 s. Aggregate evaluation results are reported in Table 1.

**Table 1. Overall Experimental Results**

| Measure | Mean |
|---|---|
| Original word count | 1,074.17 |
| Summary word count | 478.42 |
| Compression ratio | 50.83% |
| Estimated reading-time saving | 2.98 min |
| ROUGE-1 | 0.4917 |
| ROUGE-2 | 0.2371 |
| ROUGE-L | 0.2892 |
| BERTScore F1 | 0.8264 |
| LLM-as-a-Judge | 7.96/10 |
| Summary-generation time | 194.33 s |

Because the complete source rather than a human-written summary served as the reference, ROUGE and BERTScore are interpreted here as source-summary retention and similarity measures rather than conventional gold-reference performance.

### 4.2 Compression and Generation Time

Figure 2 shows the variation in compression ratio and summary-generation time across the samples. Compression ranged from 10.96% to 84.24%, indicating considerable variation in the degree of text reduction.

Generation time also varied across samples. The Pearson correlation between source length and generation time was +0.538, whereas the Spearman correlation was −0.007. The disagreement suggests that source length alone does not adequately explain the observed runtime variation.

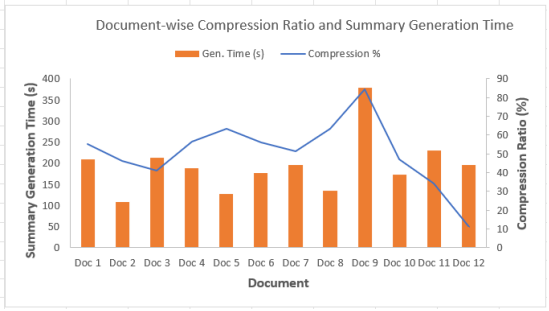


**Fig. 2.** Sample-wise comparison of compression ratio and summary-generation time.

The pipeline successfully completed local inference on the tested consumer hardware; however, the observed latency remains an important consideration for interactive use. Because runtime values were obtained from individual experimental executions rather than controlled repeated benchmark trials, they should be interpreted as observed runtimes under the tested configuration.

### 4.3 Lexical and Semantic Evaluation

Figure 3 compares ROUGE-1, ROUGE-2, and ROUGE-L across the experimental samples. ROUGE-1 was consistently higher than ROUGE-2 and ROUGE-L, with mean values of 0.4917, 0.2371, and 0.2892, respectively. The lower ROUGE-2 scores indicate lower bigram overlap than unigram overlap, which may reflect reorganization or paraphrasing of source content.

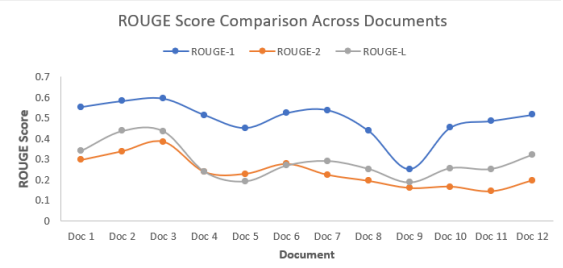


**Fig. 3.** Sample-wise comparison of ROUGE-1, ROUGE-2, and ROUGE-L scores.

BERTScore F1 averaged 0.8264 and showed less variation across samples than the ROUGE measures. Within the source-summary evaluation design, this indicates comparatively stable contextual similarity despite variation in lexical overlap. BERTScore should not, however, be interpreted as independent

evidence of factual correctness or pedagogical quality.

Figure 4 illustrates the relationship between compression and ROUGE-1. Compression was negatively associated with ROUGE-1 under both Pearson and Spearman analyses.

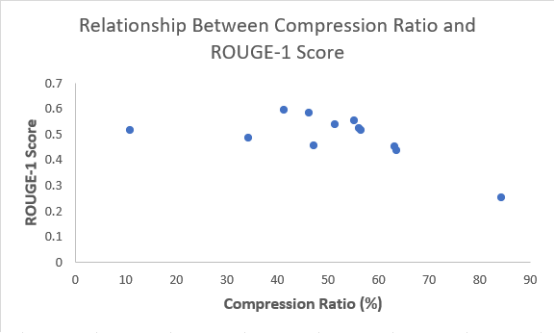


**Fig. 4.** Relationship between compression ratio and ROUGE-1 score.

This pattern should be interpreted in the context of the evaluation design. Because the complete source served as the reference, stronger compression leaves fewer source lexical units available for overlap. The observed relationship therefore represents a compression–lexical-retention pattern within the source-reference setup, rather than evidence that greater compression necessarily produces lower-quality summaries.

### 4.4 Correlation Analysis

Exploratory Pearson and Spearman correlations were calculated among selected experimental variables, as summarized in Table 2.

**Table 2. Descriptive Correlation Analysis**

| Variable Pair | Pearson | Spearman |
|---|---|---|
| Original Words ↔ Generation Time | +0.538 | −0.007 |
| Compression ↔ ROUGE-1 | −0.603 | −0.587 |
| Compression ↔ ROUGE-L | −0.495 | −0.699 |
| Compression ↔ BERTScore F1 | +0.426 | +0.378 |
| BERTScore F1 ↔ LLM Judge | +0.565 | +0.597 |

Compression was negatively associated with both ROUGE-1 and ROUGE-L, supporting the lexical-retention pattern observed in Fig. 4. In contrast, no corresponding negative association was observed between compression and BERTScore F1 within this sample. This difference illustrates why lexical and contextual similarity measures should not be treated as interchangeable.

BERTScore F1 and LLM-as-a-Judge scores were positively associated, indicating some agreement between the two evaluation signals. However, because the same Qwen3:1.7B configuration generated and judged the summaries, the LLM-based scores are not independent assessments and should be interpreted cautiously.

### 4.5 Discussion

The experiment demonstrates considerable variation in reduction behavior across the evaluated samples. Although the generated summaries were considerably shorter than their sources on average, estimated reading-time savings represent efficiency-related estimates and do not demonstrate improved comprehension, learning, or student performance.

The results also reveal different behavior between lexical and contextual evaluation measures. Reductions in source lexical overlap associated with greater compression were not accompanied by the same pattern in BERTScore. Given the source-reference design, these measures characterize different aspects of source-summary retention and similarity rather than complete summary quality or factual correctness.

The model-assisted judge provided an additional evaluation perspective and showed a positive association with BERTScore. However, because the same Qwen3 configuration generated and judged the summaries, these scores may be affected by self-evaluation bias and should remain supplementary to the automatic metrics.

Finally, the tested configuration successfully performed summarization locally on the evaluated consumer hardware. The measured runtime nevertheless identifies inference latency as an important characteristic of the local configuration. Overall, the findings support the feasibility of the tested local summarization pipeline while highlighting the importance of interpreting compression, lexical similarity, contextual similarity, model-assisted evaluation, and runtime as complementary rather than interchangeable measures.

## 5. LIMITATIONS

This study has several limitations. First, the evaluation was conducted on a relatively small dataset of 12 manually selected English-language educational samples covering data mining, machine learning, artificial intelligence, and knowledge-based systems. The dataset does not represent the broader diversity of educational materials across disciplines, languages, document structures, and levels of academic complexity. Manual selection may also introduce selection bias; therefore, the findings should be interpreted as characteristics of the evaluated dataset rather than broadly generalizable performance estimates.

Second, the experiment evaluated a single locally deployed Qwen3:1.7B configuration through Ollama, without direct comparison with other local models, larger models, cloud-based systems, or conventional summarization approaches. The study therefore characterizes the feasibility and behavior of the tested configuration rather than establishing its superiority. In addition, the pipeline used fixed, non-overlapping approximately 1,000-word chunks followed by consolidation. Alternative chunk sizes, overlapping strategies, and consolidation methods were not evaluated and may influence information continuity and retention.

Third, the evaluation metrics have methodological constraints. ROUGE and BERTScore used the complete source as the reference rather than an independently prepared human summary. Consequently, they are interpreted as source-summary lexical retention and contextual similarity measures rather than conventional reference-summary performance or independent evidence of factual correctness and pedagogical quality. The LLM-as-a-Judge assessment also used the same Qwen3:1.7B configuration for generation and evaluation, creating potential self-evaluation bias. Its scores are therefore treated as supplementary model-assisted assessments rather than independent ground truth.

Finally, runtime values were obtained from individual experimental executions rather than controlled repeated benchmarking trials. Hardware utilization, memory and energy consumption, and performance across different computing environments were not systematically evaluated. The study also did not include student or instructor evaluation or measure learning outcomes such as comprehension, knowledge retention, examination performance, or perceived usefulness. Accordingly, the observed text reduction and estimated reading-time savings do not establish improvements in educational outcomes.

Taken together, these limitations define the study as an exploratory evaluation of a locally deployed Qwen3 configuration for educational document summarization. Future work should extend the evaluation to larger and multilingual datasets, multiple model configurations, independently prepared reference summaries, human evaluation, alternative chunking strategies, and systematic computational benchmarking.

## 6. CONCLUSION AND FUTURE WORK

This study evaluated EduAssist as an experimental framework for locally deployed educational document summarization using Qwen3:1.7B through Ollama. Across 12 English-language educational samples, the model achieved a mean compression ratio of 50.83% and an estimated reading-time saving of 2.98 min. Mean BERTScore F1 was 0.8264, the mean LLM-as-a-Judge score was 7.96/10, and average summary-generation time was 194.33 s per sample.

The results showed that greater compression was associated with lower source-reference lexical overlap, whereas no corresponding negative association was observed between compression and BERTScore F1 within the evaluated dataset. This difference highlights the value of complementary lexical, contextual-similarity, and model-assisted measures when examining generated summaries. Overall, the findings provide exploratory evidence for the feasibility of the evaluated Qwen3:1.7B configuration for local educational document summarization, while not establishing superiority over alternative systems, independently verified factual correctness, human-perceived quality, or improvements in learning outcomes.

Future work should evaluate larger and multilingual educational datasets, compare multiple local and cloud-based models, model sizes, and quantization configurations, and investigate alternative chunking and consolidation strategies. Evaluation should also incorporate independently prepared reference summaries and human assessment by students and educators to examine factual correctness, readability, concept retention, usefulness, and learning outcomes. Repeated runtime experiments and measurements of CPU/GPU utilization, memory consumption, and energy requirements would further enable more rigorous comparison of local deployment efficiency.

## REFERENCES

[1] Y. Zhong and D. Litman, "From Information to Insight: Leveraging LLMs for Open Aspect-Based Educational Summarization," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025, pp. 1914–1947, doi: 10.18653/v1/2025.acl-long.95.

[2] T. Xie, Y. Kuang, Y. Tang, J. Liao, and Y. Yang, "Using LLM-supported lecture summarization system to improve knowledge recall and student satisfaction," *Expert Systems with Applications*, vol. 269, Art. no. 126371, 2025, doi: 10.1016/j.eswa.2024.126371.

[3] E. D. López Zapata, C. Tang, V. Švábenský, F. Okubo, and A. Shimada, "LECTOR: Summarizing E-book Reading Content for Personalized Student Support," *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 2495–2533, 2025, doi: 10.1007/s40593-025-00478-6.

[4] S. Masri, Y. Raddad, F. Khandaqji, H. I. Ashqar, and M. Elhenawy, "Transformer Models in Education: Summarizing Science Textbooks with AraBART, MT5, AraT5, and mBART," in *Intelligent Systems, Blockchain, and Communication Technologies*, vol. 1268, Springer Nature Switzerland, 2025, pp. 286–300, doi: 10.1007/978-3-031-82377-0_25.

[5] Z. Yu, S. Liu, P. Denny, A. Bergen, and M. Liut, "Integrating Small Language Models with Retrieval-Augmented Generation in Computing Education: Key Takeaways, Setup, and Practical Insights," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education (SIGCSE TS 2025)*, 2025, pp. 1302–1308, doi: 10.1145/3641554.3701844.

[6] C. D. Jaldi, A. Saini, S. Zhang, N. Schroeder, C. Shimizu, and E. Ilkou, "Small, Private Language Models as Teammates for Educational Assessment Design," *arXiv preprint arXiv:2605.15015*, 2026.

[7] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating Summarization Evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021, doi: 10.1162/tacl_a_00373.

[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations (ICLR)*, 2020.

[9] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, A. Wang, and P. Gallinari, "QuestEval: Summarization Asks for Fact-based Evaluation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6594–6604, doi: 10.18653/v1/2021.emnlp-main.529.

[10] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, and C. Raffel, "Evaluating the Factual Consistency of Large Language Models Through News Summarization," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023, pp. 5220–5255, doi: 10.18653/v1/2023.findings-acl.322.

[11] Y. Zhong and D. Litman, "A Tale of Evaluating Factual Consistency: Case Study on Long Document Summarization Evaluation," in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 2025, pp. 12511–12532, doi: 10.18653/v1/2025.findings-acl.648.

[12] D. Li *et al*., "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China: Association for Computational Linguistics, 2025, pp. 2757–2791, doi: 10.18653/v1/2025.emnlp-main.138.