Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 1, February, Pp. 173-182

Comparative evaluation of BM25–FAISS and small-LLM–GPT in retrieval-augmented generation concept map assessments

Maskur[1,2], Didik Dwi Prasetya[*1], Triyanna Widiyaningtyas[1], Azlan Mohd Zain[3]
Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia[1]
Departement of Business Administration, State Polytechnic of Malang, Indonesia[2]
Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia[3]

Article Info
Keywords:
Retrieval-Augmented Generation, BM25, FAISS, Small-LLM, GPT, Concept Map, Assessment

Article history:
Received: November 23, 2025
Accepted: January 16, 2026
Published: February 01, 2026

Cite:
M. Maskur, D. D. Prasetya, T. Widiyaningtyas, and A. M. Zain, "Comparative Evaluation of BM25–FAISS and Small-LLM–GPT in Retrieval-Augmented Generation Concept Map Assessment", *KINETIK*, vol. 11, no. 1, Feb. 2026.
https://doi.org/10.22219/kinetik.v11i1.2594

*Corresponding author.
Didik Dwi Prasetya
E-mail address:
didikdwi@um.ac.id

Abstract
Concept map-based assessment is a practical approach to measure students' conceptual understanding, but manual assessment still faces challenges such as subjectivity, inconsistency, and limited scalability. This study proposes the application of Retrieval-Augmented Generation (RAG) as an artificial intelligence-based automated assessment solution in an educational context. The objectives of this study are to compare the effectiveness of two retrieval methods, BM25 and FAISS, and to analyse the trade-off between large-scale generative models (GPT) and Small-LLM in assessing concept map propositions. This study uses a quantitative experimental approach by combining a retriever and a generator in the RAG system. Performance evaluation is carried out using the Macro-F1 and QWK metrics to measure agreement with expert judgment, and the Explanation Relevance Score (ERS) to assess explanation quality. The experimental results show that the FAISS–GPT combination achieves the best performance, with a Macro-F1 of 0.338 and a QWK of 0.146, slightly superior to the BM25–GPT combination. In contrast, the use of Small-LLM, both with BM25 and FAISS, showed lower performance with Macro-F1 values in the range of 0.167–0.221 and QWK close to zero. This finding confirms that semantic-based retrieval plays a vital role in improving the accuracy of automated assessment, while large-scale generative models are more effective in representing conceptual relationships in depth. This study contributes through a comparative analysis of retrievers and generators, and by introducing ERS as an additional metric for RAG-based automated assessment in the field of education.

1. Introduction

Artificial Intelligence (AI) has rapidly transformed various domains, including education. AI capabilities now extend beyond basic computations to include language comprehension, pattern recognition, and the provision of actionable recommendations. In educational settings, AI facilitates adaptive learning, learning analytics, and automated assessments, thereby enhancing the efficiency, fairness, and consistency of evaluations [1][2]. As educational paradigms shift toward skills and learning outcomes, assessments are increasingly expected to capture not only scores but also the depth of students' conceptual understanding [3][4]. Consequently, methods that illustrate relationships between ideas have gained prominence. Concept maps, which represent knowledge through interconnected concepts, are widely utilized for this purpose [5][6]. They are effective in evaluating higher-order thinking, the integration of ideas, and overall comprehension. Large Language Models (LLMs), such as GPT, have significantly expanded AI's capabilities for processing language [7][8][9]. These models excel at contextual understanding, generating clear explanations, and reasoning through complex responses. Due to these strengths, LLMs are particularly promising for automated assessments that require both feedback and explanatory support, rather than merely providing scores [10].

To make generative models more reliable and grounded in facts, the Retrieval-Augmented Generation (RAG) approach was developed. RAG combines a search tool that finds useful external information with a generator that uses it to create answers [11][12][13]. This setup makes AI responses more accurate, clear, and easy to check, and is now a leading method for systems that need both understanding and trustworthy information. In educational contexts, RAG is increasingly employed in assessment systems that prioritize conceptual understanding. By combining retrieval and generation, RAG enables AI to evaluate not only final answers but also the relationships among concepts articulated by students. This capability makes RAG particularly suitable for concept map assessments, where learning quality is determined by the strength of conceptual linkages [14].

Even though there is more research on AI in educational assessment, using RAG for concept map assessment is still rare. Most studies focus on simple text answers rather than the relationship analysis that concept maps require. As a result, RAG's full potential for structured concept assessment has not been fully explored [15][16]. Additionally,

few studies have compared retrieval methods such as Best Matching 25 (BM25) and Facebook AI Similarity Search (FAISS) within educational contexts [17][18]. Although these methods employ distinct strategies for identifying relevant information, their impact on concept map assessment outcomes remains unclear. Large models like GPT are good at understanding meaning but require substantial computing power. Smaller models (Small-LLMs) use less power but often have trouble with complex ideas [19][20]. So, finding the right balance between accuracy and efficiency in educational RAG systems is still an [21]. Most prior studies evaluate system performance primarily using quantitative metrics such as Macro-F1 and Quadratic Weighted Kappa (QWK), with limited attention to the quality of generated explanations [22]. However, in educational settings, clear and relevant explanations are essential for both comprehension and instruction. Without assessing explanation quality, automated assessments may achieve accuracy but fail to support meaningful learning. These gaps show that more research is needed to fully evaluate RAG-based concept map assessment systems, especially in combining different retrievers and generators, balancing efficiency and accuracy, and assessing explanation quality [23].

To address these gaps, this study proposes a comparative evaluation framework based on Retrieval-Augmented Generation (RAG) for automated assessment of concept map propositions. The framework combines two retrievals. To fill these gaps, this study introduces a framework that uses Retrieval-Augmented Generation (RAG) to assess concept map statements automatically. The framework tests two retrieval methods, BM25 and FAISS, with two generators, GPT and a Small-LLM, to compare how well they work, how efficient they are, and how good their explanations are. The framework also uses a metric, the Explanation Relevance Score (ERS), which measures the semantic alignment between system-generated explanations [24][25]. By integrating both quantitative and qualitative evaluations, this study offers a more comprehensive assessment. The primary contributions of this research are threefold. First, it provides a systematic comparison of lexical retrieval (BM25) and semantic retrieval (FAISS) in the context of concept map-based assessment, addressing a gap in the existing literature. Second, it provides an empirical evaluation of the trade-off between performance and efficiency when employing large-scale models (GPT) versus lightweight models (Small-LLM) as generators in RAG systems. Third, it introduces the Explanation Relevance Score (ERS) as a novel metric for evaluating explanation quality semantically and contextually.

By integrating semantic retrieval, generative reasoning, and explanation-based evaluation, this research advances the development of automated assessment systems that are accurate, transparent, and pedagogically valuable. The proposed framework seeks to enhance the fairness, trustworthiness, and relevance of educational assessments in contemporary AI-driven learning environments.

2. Research Method

We used a comparative experimental approach to test how well two information retrieval methods, BM25 and FAISS, work with two generative models, Small-LLM and GPT, for assessing propositions in concept maps. BM25 is a keyword-matching algorithm chosen for its reliable performance in tasks that require accurate word matching. It measures term relevance based on word frequency and keyword matches, making it well-suited for applications that need precision, but it does not capture the meaning of concepts. FAISS, on the other hand, uses semantic, vector-based retrieval by turning text into numerical embeddings to reflect deeper context and relationships. For generating responses, GPT provides detailed, context-based explanations, while Small-LLM offers faster, more efficient answers with more straightforward reasoning. By combining these retrieval (BM25, FAISS) and generation (GPT, Small-LLM) methods, our approach aims to improve the accuracy, consistency, and clarity of automated concept map assessments. The goal is to create a transparent and efficient AI-based evaluation system that supports meaningful educational assessment and feedback in digital learning environments [26][27]. All stages are presented in Figure 1.

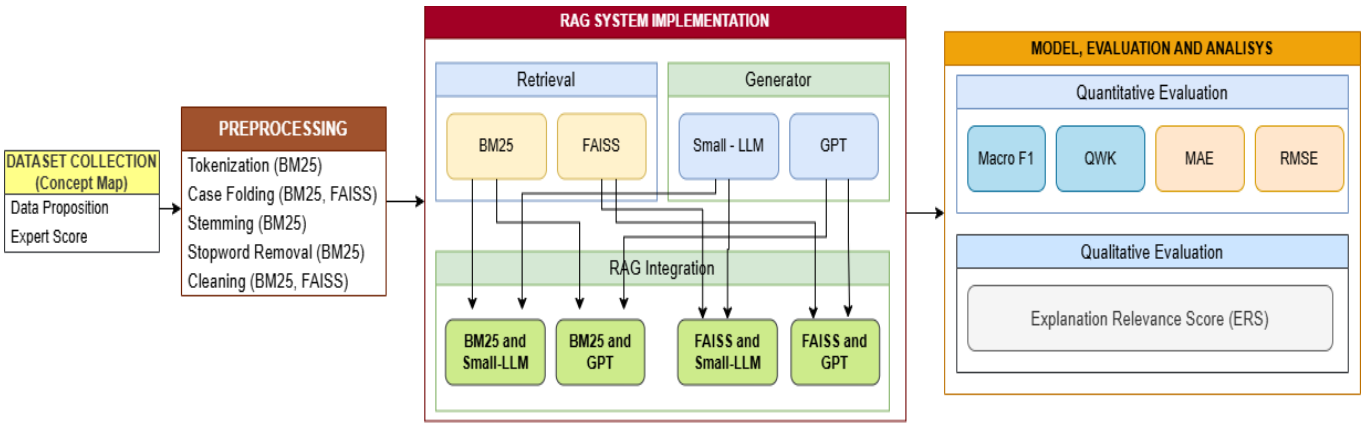

Figure 1. Research Workflow

2.1 Dataset Collection

The dataset consists of 691 propositions, which are student answers and scores assessed by a senior and experienced teacher who has taught for 11 years teaching the control class and the experimental class with a Scale (0–3) that is 0: Incorrect; 1: Partially correct; 2: Correct but superficial; 3: Scientifically sound and well-reasoned. These scores reflect the level of truth, accuracy, and conceptual quality of each proposition used as ground truth. The dataset's domain is educational, making it relevant for automated assessment tasks [28]. Relational Database Concept Map Dataset Table 1 below.

Table 1. Relational Database Concept Map Dataset

No	Proposition	Score
1	basis data relasional keuntungan mudah melakukan operasi data	2
2	basis data relasional keuntungan sederhana	1
3	basis data relasional pengertian tabel dua dimensi	3
4	basis data relasional memiliki domain	3
5	tabel dua dimensi disebut relasi	2
6	basis data relasional memiliki cardinality	3
7	basis data relasional memiliki atribut	3
8	basis data relasional terdiri dari tupel	3
9	basis data relasional key relational key	2
10	relational key key super key	2
11	relational key key candidate key	3
12	relational key key primary key	3
13	relational key key alternate key	3
14	relational key key foreign key	3
15	relasi aturan aturan integritas	3
16	super key adalah kunci utama	3
17	primary key adalah kunci utama	3
18	foreign key adalah kunci tamu	3
19	basis data relasional memiliki bahasa	3
20	bahasa disebut SQL	0

2.2 Data Preprocessing

Text Preparation Procedures for BM25 and FAISS Information Retrieval Systems. The BM25 technique employs methods such as converting all letters to lowercase, dividing text into tokens, eliminating common words, applying stemming or lemmatisation, and sanitising the text to improve the precision of word matches. These steps minimise word form variations and eliminate superfluous textual data. Changing the text to lowercase is accomplished through case folding; tokenisation breaks the text into smaller chunks; stopword removal removes trivial words; and stemming or lemmatisation condenses different word forms. Text cleaning is executed to extract unnecessary characters. Conversely, the FAISS method places greater emphasis on understanding the significance of words via embeddings. Because this data helps analyse the full context, it does not discard stopwords or perform stemming or lemmatisation. To retain clarity, FAISS also retains important punctuation, whereas BM25 focuses more on exact word comparisons. The data preprocessing process is presented in Table 2.

Table 2. Data Preprocessing

Preprocessing Step	BM25 (Lexical-Based Retrieval)	FAISS (Vector-Based Retrieval)
Case Folding	✓ Done to uniformize all text into lowercase letters	✓ Generally applied as normalization before the embedding process.
Tokenization	✓ This is done because the BM25 method calculates the frequency and distribution of tokens.	~ Not done manually; the tokenization process is handled by the embedding model's internal tokenizer.
Stopword Removal	✓ Done to reduce common words that do not contribute to the search.	✗ Not done because stopwords help the model understand sentence structure and context.
Stemming / Lemmatization	✓ This is done to reduce the variation in word forms so that lexical matching is more precise.	✗ Not done because it can eliminate the original word form which has important semantic meaning for embedding.

Cleaning	✓ Done to remove irrelevant characters and improve token quality.	✓ Limited use; Use but important punctuation is retained to support understanding of the context

2.3 Implementasi Retrieval-Augmented Assessment

This stage is the core of the research, where the Retrieval-Augmented Generation system is applied to automatically assess concept map propositions and produce explanations that are relevant to the level of understanding being measured.

2.3.1 Retriever: BM25 dan FAISS

The retriever is tasked with finding and calculating similarities between student propositions and the knowledge base or reference propositions from experts. Two comparison methods are used.

a. BM25 (*Best Matching 25*)

BM25 functions as a retriever algorithm that generates a list of relevant documents along with their relevance scores, which are then submitted to the Generative Model (Small-LLM / GPT) to be used in generating answers [29]. BM25 is measured using Equation 1.

$$BM25(Q,D) = \sum_{t \in Q} I\,DF(t) \times \frac{f(t,D)(k_1+1)}{f(t,D) + k_1\left(1 - b + b\frac{|D|}{avgdl}\right)} \quad (1)$$

b. FAISS (*Facebook AI Similarity Search*)

FAISS acts as a retrieval component that finds the most suitable documents based on semantic similarity. These documents are then passed to a Generative Model (Small-LLM/GPT), which processes them to generate answers [30]. FAISS is measured using Equation 2.

$$\text{Cosine Similarity}(Q,D) = \frac{Q \cdot D}{||Q|| \,.\, ||D||} \quad (2)$$

c. RAG Integration Process

The student's proposition becomes a query into the system. Next, the retriever (BM25 or FAISS) searches for similarities and assigns an initial score. Then the score and proposition text are passed to the generator Small-LLM or GPT. The generator produces: (1) an automatic score (prediction), (2) a brief explanation of the reason for the score.

2.3.2 Small-LLM and GPT Generators

The generator component is responsible for generating automatic explanations of the assessment results, with two comparison models: (1) Small-LLM (Large Language Model) is a small language model that is run locally in this study using phi-3-mini. This model then builds explanations based on simple linguistic patterns. (2) GPT (Generative Pre-training Transformer) is a large model based on the Generative Transformer with deeper semantic understanding capabilities used is gpt-oss. This model then provides more contextual and specific explanations.

2.3.3 Hyperparameter Settings and Model Configuration

To ensure experimental consistency and enable equitable comparisons across combinations of retrieval techniques and generators, hyperparameter settings and model configurations were defined. The number of pertinent documents recovered (top-k) was fixed at 3 per trial, and BM25 was used for the lexical retrieval step via the rank-bm25 library, with no changes to its internal parameters. Two different language models were employed in the generation step. Phi-3-mini running locally was used to implement the Small-LLM model. The OpenRouter API was used to retrieve the GPT model utilising gpt-oss for comparison.

2.3.4 RAG Integration Process

2x2 testing scenario to see better performance and results. (1) BM25 × Small-LLM, (2) BM25 × GPT, (3) FAISS × Small-LLM, (4) FAISS × GPT. The testing stage is carried out in several stages as follows. First, the student's proposition becomes a query into the system. Second, the Retriever (BM25 or FAISS) looks for similarities and gives an initial score. Third, the score and proposition text are passed to the generator (Small-LLM or GPT). Fourth, the generator produces an automatic score and a brief explanation of the reasons for giving the score.

2.4 Model Evaluation and Analysis

This stage aims to assess the extent to which the Retrieval-Augmented Assessment (RAG) system being developed is capable of providing propositional assessments and explanations that closely align with expert assessments. The evaluation is conducted quantitatively and qualitatively, encompassing aspects of accuracy, relevance, efficiency, and validity.

2.3.1 Quantitative Evaluation

Quantitative evaluation is used to measure system performance based on a comparison of model predictions (RAG) with expert reference values. The evaluation method utilizes several statistical and machine learning metrics, including:

1. Mean Squared Error (MSE)
   Calculate the squared difference between expert scores and predictions [31]. Mean Squared Error is measured using Equation 3.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

2. Root Mean Squared Error (RMSE)
   MSE root which provides an interpretation comparable to the score [32]. Root Mean Squared Error is measured using Equation 4.

$$RMSE = \sqrt{\text{MSE}} \tag{4}$$

3. Mean Absolute Error (MAE)
   Calculate the absolute difference between expert and model scores [33]. Mean Absolute Error is measured using Equation 5.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i \, | \tag{5}$$

4. Quadratic Weighted Kappa (QWK)
   A statistical metric used to measure the level of agreement between two annotators [34]. Quadratic Weighted Kappa is measured using Equation 6.

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{i,j} \cdot O_{i,j}}{\sum_{i,j} w_{i,j} \cdot E_{i,j}} \tag{6}$$

5. F1-Score
   A measure of classification model performance, which is the harmonic mean between Precision and Recall [35]. The F1-Score is measured using Equation 7.

$$F1 = 2\text{x}\frac{Precision \; x \; Recall}{Precision + \; Recall} \tag{7}$$

6. Cosine Similarity
   To calculate the similarity to the Explanation Relevance Score (ERS) using Cosine Similarity with E = embedding of the RAG explanation and P = embedding of the proposition [36].The Cosine Similarity is measured using Equation 8.

$$cosine\_similarity(E,P) = \frac{E \cdot P}{\| \, E \, \| \cdot \| \, P \, \|} = \frac{\sum_{i=1}^{d} E_i \cdot P_i}{\sqrt{\sum_{i=1}^{d} E_i^2} \cdot \sqrt{\sum_{i=1}^{d} P_i^2}} \tag{8}$$

2.3.2 Qualitative Evaluation

Qualitative evaluation is conducted to analyze the relevance of the explanation (Explanation Relevance Score) generated by the model. Aspects assessed include the Explanation Relevance Score, which assesses the quality of the explanation generated by the model (Small-LLM and GPT), assessed from the semantic relevance between the model's explanation and the expert's reasons or criteria in giving the score.

3. Results and Discussion

This section reports the results of evaluating the Retrieval-Augmented Generation (RAG) system for automated assessment of concept map propositions. The evaluation addressed five primary areas: the performance of retrieval methods (BM25 and FAISS), the effectiveness of generator models (Small-LLM and GPT), outcomes of various retriever-generator pairings, the trade-off between accuracy and efficiency and the relationship between explanation quality and score consistency. The objective was to identify the optimal RAG configuration for accurate assessment and clear explanations. The findings indicate that both retrieval quality and generator capabilities are critical for practical concept map assessment using the RAG system.

3.1 Retrieval Method Performance (BM25 and FAISS)

The BM25 Retriever results in Table 3 show that the proposition “relational key 4 alternate key” obtained the highest score of 6.6313 because it has a firm lexical match with the query. In contrast, the lowest score of 0 appears between the propositions “domain term relational database” and “domain term relational database,” which have little term overlap. In contrast, in the FAISS Retriever in Table 4, the same proposition obtained the highest score of 1, which represents the most substantial semantic similarity, while the lowest score of 0.2306 is found in the proposition “relational database 2 simple form,” which still shows closeness of meaning even without a direct word match. This difference confirms that BM25 emphasises lexical similarity, while FAISS is more effective in capturing semantic similarity.

Table 3. Retrieval Results BM25

Proposition	Value
relational key 4 alternate key	6.6313
relational key 5 foreign key	2.9397
basisdata relasional memiliki relational key	2.2864
basis data relasional istilah domain	0
basisdata relasional istilah istilah domain	0

Table 4. Retrieval Results FAISS

Proposition	Value
relational key 4 alternate key	1
relational key 5 foreign key	0.7129
basisdata relasional memiliki relational key	0.5433
bahasa komersial adalah sql	0.2901
basis data relasional 2 bentuknya sederhana	0.2306

3.2 Model Generator Performance (Small-LLM and GPT)

Model generator performance was evaluated using Macro-F1, QWK, MAE, RMSE, ERS, and inference time metrics. Test results showed that GPT consistently outperformed Small-LLM across all retriever configurations. The FAISS–GPT combination produced the highest Macro-F1 (0.338) and QWK (0.146), with the lowest MAE and RMSE of 0.973 and 1.321, respectively. Conversely, BM25–Small-LLM performed the lowest, with a Macro-F1 of 0.167 and a negative QWK (-0.014), indicating a poor match between the scores and the reference assessment. In terms of explanation quality, GPT achieved the highest ERS in the BM25–GPT combination (0.920), while FAISS–Small-LLM also demonstrated a relatively high ERS (0.880). This indicates that narrative quality does not always align with the consistency of the numerical scores. Qualitatively, Small-LLM produces deterministic and uniform score-based explanations, thus functioning as a rule-based explanation. In contrast, GPT produces more contextually rich explanations and integrates domain knowledge as a semantic reasoning generator. In terms of efficiency, Small-LLM is significantly faster, with an inference time of under 0.02 seconds, while GPT requires approximately 7.5 seconds.

This finding confirms the trade-off between semantic quality and computational efficiency in RAG systems. The results of the Retrieval-Augmented Assessment are presented in Table 5.

Table 5. Retrieval-Augmented Assessment Results

Retriever	Generator	Proposition	Expert Score	Score Prediction	Explanation
BM25	GPT	bahasa disebut sql	0	2	Score: 2 Explanation: Konsep "SQL" sebagai bahasa dimengerti secara umum, tetapi tidak menggambarkan dengan ketepatan bahwa SQL adalah bahasa pendzijurik khusus untuk interaksi dengan basis data relasional. Istilah "bahasa" terlalu luas, dan proposisi ini mengurangi kompleksitas konteks (misalnya, SQL adalah satu dari banyak bahasa database, bukan tanpa koiniczny jenis). Relevansi ada, tetapi terminologi kurang lengkap. (time: 7.15s)
BM25	Small-LLM	bahasa disebut sql	0	1	Proposisi 'bahasa disebut sql' dinilai 1 oleh Small-LLM.
FAISS	GPT	bahasa disebut sql	0	0	Score: 0 Explanation: Pernyataan "bahasa disebut sql" tidak akurat. SQL adalah akronim dari Structured Query Language (Bahasa Percakapan Terstruktur), yang merujuk pada bahasa pemrograman khusus untuk memanipulasi data dalam basis data relasional. Dinyatakan bahwa istilah "bahasa" (yang merujuk pada konsep umum) "disebut SQL" tidak sesuai, karena SQL adalah bahasa spesifik, bukan nama umum untuk semua bahasa pemrograman. (time: 7.77s)
FAISS	Small-LLM	bahasa disebut sql	0	0	Proposisi 'bahasa disebut sql' dinilai 0 oleh Small-LLM.

3.3 Evaluation of the Retriever and Generator Combination

A retriever-generator combination evaluation was conducted to assess the performance of the RAG system as an integrated architecture. Four configurations were tested: BM25–GPT, BM25–Small-LLM, FAISS–GPT, and FAISS–Small-LLM. The test results showed that the FAISS–GPT configuration provided the best performance with a Macro-F1 score of 0.338, a QWK of 0.146, and the lowest MAE and RMSE of 0.973 and 1.321, respectively. These findings indicate that integrating semantic-based retrieval with a high-capacity generator provides the most consistent and accurate assessment results. Conversely, the BM25–Small-LLM combination produced the lowest performance with a Macro-F1 score of 0.167 and a negative QWK (-0.014), indicating a discrepancy between the scores and the reference assessment. The FAISS–Small-LLM combination demonstrated improvement over BM25–Small-LLM, but remained below the GPT-based configuration. These results confirm that the performance of an RAG system is not solely determined by the quality of the generator model, but is also significantly influenced by the retrieval method used. Thus, the success of a RAG system in assessing concept map propositions depends on the synergy between semantic retrieval and the generative model's reasoning capabilities. The results of the LLM Model Evaluation on the BM25 and FAISS Methods using Macro F1, QWK and MAE are presented in Figure 2.

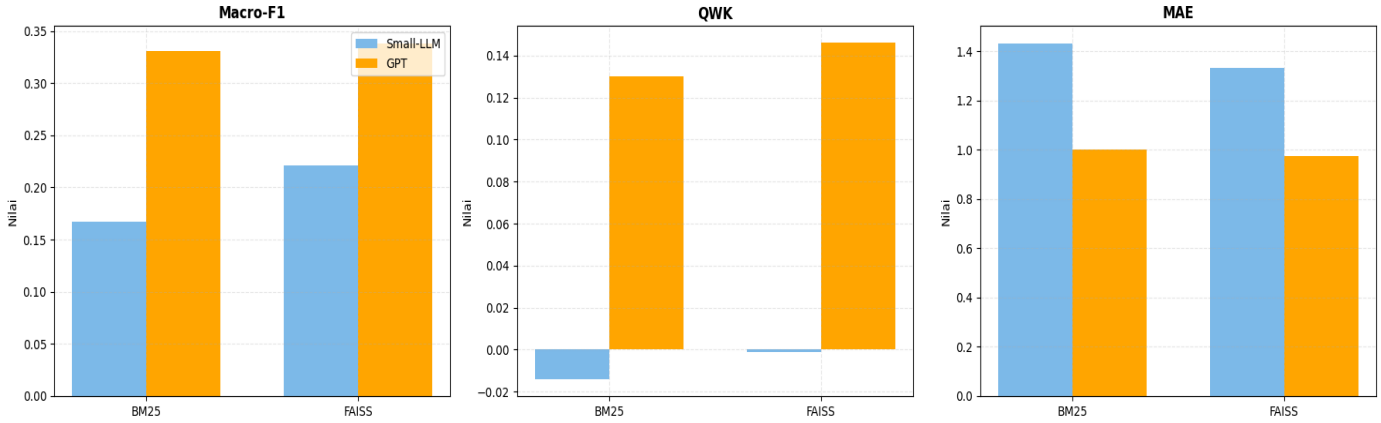

Figure 2. Comparison chart of LLM Model Evaluation Results on BM25 and FAISS Methods

3.4 Macro-F1 Performance and Efficiency Trade-off

Table 6 shows that the GPT-based system achieved the highest Macro-F1, but its inference time exceeded 7 seconds. In contrast, Small-LLM responded almost instantly but was much less accurate. These findings suggest that the configuration choice should align with the application's needs.

Table 6. Macro-F1 Performance and Efficiency Results

Configuration	Macro-F1	Time (s)
FAISS–GPT	0.338	7.488
BM25–GPT	0.331	7.746
FAISS–Small-LLM	0.221	0.015
BM25–Small-LLM	0.167	0.001

3.5 Explanation Quality and Score Consistency

Table 7 shows that BM25–GPT achieves the highest ERS (0.920), but its QWK is lower than that of FAISS–GPT. This means that strong explanations do not always lead to consistent scores. These results point to a gap between generative abilities and ordinal scoring methods.

Table 7. Explanation Quality and Consistency Score Results

Combination	ERS	QWK
BM25–GPT	0.920	0.130
FAISS–GPT	0.790	0.146
Small-LLM based	≤0.880	≤0

The experimental results indicate that the FAISS–GPT configuration achieves the best performance in assessing concept map propositions, with a Macro-F1 of 0.338 and a QWK of 0.146, slightly outperforming BM25–GPT (Macro-F1 = 0.331; QWK = 0.130). This finding confirms that dense semantic retrieval provides more relevant contextual information than lexical-based retrieval, enabling the generative model to produce assessments closer to expert judgment. In contrast, configurations involving Small-LLM show a substantial performance decline, with Macro-F1 values ranging from 0.167 to 0.221 and QWK values near zero or negative. This result supports previous studies reporting that lightweight generative models have limited capability to replicate expert-level assessment, particularly for ordinal scoring tasks that require sensitivity to subtle conceptual differences. Although GPT achieves a high Explanation Relevance Score (ERS = 0.920), the agreement between predicted and expert scores remains moderate, indicating that convincing explanations do not always guarantee numerical accuracy. Similar inconsistencies between explanation quality and scoring reliability have also been reported in earlier RAG and generative assessment studies.

Compared with prior RAG research by Fukui et al. and Xu et al., which mainly focused on question answering and medical domains [37], [38]. This study extends RAG evaluation to educational concept map assessment and explicitly distinguishes between sparse and dense retrieval strategies. Moreover, consistent with Prasetya et al., the results highlight the importance of deep conceptual understanding and further demonstrate that retrieval type and generator capacity must be jointly optimized [39].

Finally, the average GPT inference time of approximately 7.5 seconds illustrates a clear trade-off between accuracy and efficiency. Overall, the findings confirm that effective RAG-based assessment systems require balanced integration of retrieval and generation components to ensure both reliability and practical applicability.

4. Conclusion

Based on a study on the application of Retrieval-Augmented Generation (RAG) for automated concept map proposition assessment, two retrieval methods (BM25 and FAISS) and two generator models (Small-LLM and GPT) were compared. The test results showed that the FAISS–GPT configuration provided the best performance, with a Macro-F1 value of 0.338 and a QWK of 0.146, and the lowest MAE and RMSE values of 0.973 and 1.321, respectively. This performance was slightly higher than that of BM25–GPT (Macro-F1 = 0.331; QWK = 0.130), confirming that semantic-based retrieval provides more relevant context for the generative model. In contrast, the Small-LLM-based configuration showed a significant performance decline, with Macro-F1 ranging from 0.167 to 0.221 and QWK approaching zero or even negative. Nevertheless, Small-LLM excelled in efficiency, with an inference time of under 0.02 seconds, compared to GPT, which required approximately 7.5 seconds per proposition. This finding suggests a clear trade-off between assessment accuracy and computational efficiency. In terms of explanation quality, GPT achieved the highest Explanation Relevance Score (ERS) of 0.920. However, this result was not always accompanied by consistent numerical scores, indicating that the quality of the explanation narrative does not always directly impact assessment accuracy. The main contribution of this research lies in the comprehensive evaluation of the RAG system in the context of concept map-based assessment, confirming that the system's success depends heavily on the integration of strategy decision-making and the model generator's capacity. This research extends the application of RAG to the educational domain and provides an empirical basis for developing more valid and reliable automated assessment systems.

Future Work

Future research can develop more robust retrieval techniques and test the system on larger, multilingual educational datasets. Small-LLM's performance also needs to be improved through fine-tuning to maintain efficiency while increasing accuracy. Analyzing the cost of input-output tokens when using GPT is crucial for understanding the trade-off between accuracy, explanation quality and cost-efficiency. Furthermore, strategies such as iterative retrieval and fact-checking are needed to reduce hallucinations and improve the consistency of RAG systems.

Acknowledgement

The first author would like to express his gratitude to the Ministry of Higher Education, Science, and Technology (Kemendiktisaintek) for Publication Assistance in Reputable Journal in 2025, according to the letter number: 0488/C/DT.06.01/2025

References

[1] S. Bouguettaya, F. Pupo, M. Chen, and G. Fortino, "A Meta-Survey of Generative AI in Education: Trends, Challenges, and Research Directions," Sep. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: https://doi.org/10.3390/bdcc9090237

[2] C. Cohn, N. Hutchins, T. Le, and G. Biswas, "A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23182–23190, 2024, doi: https://doi.org/10.1609/aaai.v38i21.30364

[3] S. Y. Yanes-Luis, D. G. Gutiérrez-Reina, and S. T. Marín, "Towards a Retrieval-Augmented Generation Framework for Originality Evaluation in Projects-Based Learning Classrooms," *Educ. Sci. (Basel).*, vol. 15, no. 6, 2025, doi: https://doi.org/10.3390/educsci15060706

[4] R. Waterman, B. Lafving, C. Okar, and N. Jain, "A Custom GPT for Executive MBA Students: A Case Study in Enhancing Learning," *Stat*, vol. 14, no. 4, 2025, doi: https://doi.org/10.1002/sta4.70109

[5] T. Evans and I. Jeong, "Concept maps as assessment for learning in university mathematics," *Educational Studies in Mathematics*, vol. 113, no. 3, pp. 475–498, 2023, doi: https://doi.org/10.1007/s10649-023-10209-0

[6] K. E. de Ries, H. Schaap, A.-M. M. J. A. P. van Loon, M. M. H. Kral, and P. C. Meijer, "A literature review of open-ended concept maps as a research instrument to study knowledge and learning," *Qual. Quant.*, vol. 56, no. 1, pp. 73–107, Feb. 2022, doi: https://doi.org/10.1007/s11135-021-01113-x

[7] P. Dahal, S. Nugroho, C. Schmidt, and V. Sanger, "AI-Based Learning Recommendations: Use in Higher Education †," *Future Internet*, vol. 17, no. 7, 2025, doi: https://doi.org/10.3390/fi17070285

[8] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, "An LLM-Driven Chatbot in Higher Education for Databases and Information Systems," *IEEE Transactions on Education*, vol. 68, no. 1, pp. 103–116, 2025, doi: https://doi.org/10.1109/TE.2024.3467912

[9] D. Hennekeuser, D. D. Vaziri, D. Golchinfar, D. Schreiber, and G. Stevens, "Enlarged Education – Exploring the Use of Generative AI to Support Lecturing in Higher Education," *Int. J. Artif. Intell. Educ.*, vol. 35, no. 3, pp. 1096–1128, 2025, doi: https://doi.org/10.1007/s40593-024-00424-y

[10] P. Fergus *et al.*, "Towards Context-Rich Automated Biodiversity Assessments: Deriving AI-Powered Insights from Camera Trap Data," *Sensors*, vol. 24, no. 24, 2024, doi: https://doi.org/10.3390/s24248122

[11] M. Klesel and H. F. Wittmann, "Retrieval-Augmented Generation (RAG)," *Business & Information Systems Engineering*, vol. 67, no. 4, pp. 551–561, 2025, doi: https://doi.org/10.1007/s12599-025-00945-3

[12] B. E. Perron, B. S. Hiltz, E. M. Khang, and S. A. Savas, "AI-Enhanced Social Work: Developing and Evaluating Retrieval-Augmented Generation (RAG) Support Systems," *J. Soc. Work Educ.*, vol. 61, no. 1, pp. 3–13, 2025, doi: https://doi.org/10.1080/10437797.2024.2411172

[13] Y. Lee, "Developing a computer-based tutor utilizing Generative Artificial Intelligence (GAI) and Retrieval-Augmented Generation (RAG)," *Educ. Inf. Technol. (Dordr).*, vol. 30, no. 6, pp. 7841–7862, 2025, doi: https://doi.org/10.1007/s10639-024-13129-5

[14] C. Cole, A. Hajikhani, E. Hylkilä, E. Paronen, and H. Pihkola, "Towards AI-augmented sustainability assessments: integrating large language models in the case of product social life cycle assessment," *International Journal of Life Cycle Assessment*, 2025, doi: https://doi.org/10.1007/s11367-025-02508-w
[15] F. Noorbehbahani and A. A. Kardan, "The automatic assessment of free text answers using a modified BLEU algorithm," *Comput. Educ.*, vol. 56, no. 2, pp. 337–345, Feb. 2011, doi: https://doi.org/10.1016/j.compedu.2010.07.013
[16] W.-J. HOU and J.-H. TSAO, "AUTOMATIC ASSESSMENT OF STUDENTS' FREE-TEXT ANSWERS WITH DIFFERENT LEVELS," *International Journal on Artificial Intelligence Tools*, vol. 20, no. 02, pp. 327–347, Apr. 2011, doi: https://doi.org/10.1142/S0218213011000188
[17] L. D. Krisnawati, A. W. Mahastama, S.-C. Haw, K.-W. Ng, and P. Naveen, "Indonesian-English Textual Similarity Detection Using Universal Sentence Encoder (USE) and Facebook AI Similarity Search (FAISS)," *CommIT (Communication and Information Technology) Journal*, vol. 18, no. 2, pp. 183–195, Sep. 2024, doi: https://doi.org/10.21512/commit.v18i2.11274
[18] G. Dobriţa, S.-V. Oprea, and A. Bâra, "An NLP-driven e-learning platform with LLMs and graph databases for personalised guidance," *Conn. Sci.*, vol. 37, no. 1, Dec. 2025, doi: https://doi.org/10.1080/09540091.2025.2518991
[19] J. B. Vargas Bernuy, M. A. Nolasco-Mamani, N. C. Velásquez Rodríguez, R. L. Gambetta Quelopana, A. N. Martinez Valdivia, and S. M. Espinoza Vidaurre, "Relative Advantage and Compatibility as Drivers of ChatGPT Adoption in Latin American Higher Education: A PLS SEM Study Towards Sustainable Digital Education," *Sustainability*, vol. 17, no. 18, p. 8329, Sep. 2025, doi: https://doi.org/10.3390/su17188329
[20] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets," *IEEE Access*, vol. 13, pp. 94009–94025, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3574629
[21] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 6769–6781. doi: https://doi.org/10.18653/v1/2020.emnlp-main.550
[22] S. Aksoy and A. Daou, "An Explainable Web-Based Diagnostic System for Alzheimer's Disease Using XRAI and Deep Learning on Brain MRI," *Diagnostics*, vol. 15, no. 20, p. 2559, Oct. 2025, doi: https://doi.org/10.3390/diagnostics15202559
[23] D. de Oliveira, "Assessing ChatGPT in digital education: A case study on student perception," *Digital Engineering*, vol. 7, p. 100067, Dec. 2025, doi: https://doi.org/10.1016/j.dte.2025.100067
[24] J. S. Jauhiainen and A. G. Guerra, "Evaluating Students' Open-ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large," *Advances in Artificial Intelligence and Machine Learning*, vol. 4, no. 4, pp. 3097–3113, 2024, doi: https://doi.org/10.54364/AAIML.2024.44177
[25] F. Fanelli, M. Saleh, P. Santamaria, K. Zhurakivska, L. Nibali, and G. Troiano, "Development and Comparative Evaluation of a Reinstructed GPT-4o Model Specialized in Periodontology," *J. Clin. Periodontol.*, vol. 52, no. 5, pp. 707–716, 2025, doi: https://doi.org/10.1111/jcpe.14101
[26] S. Elmitwalli, J. Mehegan, S. Braznell, and A. Gallagher, "Scalable evaluation framework for retrieval augmented generation in tobacco research using large Language models," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: https://doi.org/10.1038/s41598-025-05726-2
[27] J. Lee, S. Ahn, D. Kim, and D. Kim, "Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval," *Autom. Constr.*, vol. 168, p. 105846, Dec. 2024, doi: https://doi.org/10.1016/j.autcon.2024.105846
[28] D. D. Prasetya, A. Pinandito, Y. Hayashi, and T. Hirashima, "Analysis of quality of knowledge structure and students' perceptions in extension concept mapping," *Res. Pract. Technol. Enhanc. Learn.*, vol. 17, no. 1, p. 14, 2022, doi: https://doi.org/10.1186/s41039-022-00189-9
[29] Q. Chen, W. Zhou, J. Cheng, and J. Yang, "An Enhanced Retrieval Scheme for a Large Language Model with a Joint Strategy of Probabilistic Relevance and Semantic Association in the Vertical Domain," *Applied Sciences*, vol. 14, no. 24, p. 11529, Dec. 2024, doi: https://doi.org/10.3390/app142411529
[30] S. Xu, Z. Yan, C. Dai, and F. Wu, "MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health," *Front. Public Health*, vol. 13, Oct. 2025, doi: https://doi.org/10.3389/fpubh.2025.1635381
[31] M. Ramesh *et al.*, "Assessing WildfireGPT: a comparative analysis of AI models for quantitative wildfire spread prediction," *Natural Hazards*, vol. 121, no. 11, pp. 13117–13130, Jun. 2025, doi: https://doi.org/10.1007/s11069-025-07344-7
[32] V. Ramnarain-Seetohul, Y. Rosunally, and V. Bassoo, "A Unified Conceptual Hybrid Framework for the Automated Assessment of Short Answers," *Int. J. Artif. Intell. Educ.*, Jun. 2025, doi: https://doi.org/10.1007/s40593-025-00487-5
[33] N. Lotfy, A. Shehab, M. Elhoseny, and A. Abu-Elfetouh, "An Enhanced Automatic Arabic Essay Scoring System Based on Machine Learning Algorithms," *Computers, Materials & Continua*, vol. 77, no. 1, pp. 1227–1249, 2023, doi: https://doi.org/10.32604/cmc.2023.039185
[34] A. Doewes, N. A. Kurdhi, and A. Saxena, "Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring," 2023, doi: https://doi.org/10.5281/zenodo.8115784
[35] Y. Wang, Y. Wan, X. Lei, Q. Chen, and H. Hu, "A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models," *Array*, vol. 28, p. 100504, 2025, doi: https://doi.org/10.1016/j.array.2025.100504
[36] C. Yao and S. Fujita, "Adaptive Control of Retrieval-Augmented Generation for Large Language Models Through Reflective Tags," *Electronics (Basel).*, vol. 13, no. 23, p. 4643, Nov. 2024, doi: https://doi.org/10.3390/electronics13234643
[37] Y. Fukui, Y. Kawata, K. Kobashi, Y. Nagatani, and H. Iguchi, "Evaluation of a retrieval-augmented generation system using a Japanese Institutional Nuclear Medicine Manual and large language model-automated scoring," *Radiol. Phys. Technol.*, vol. 18, no. 3, pp. 861–876, 2025, doi: https://doi.org/10.1007/s12194-025-00941-y
[38] R. Xu, Y. Hong, F. Zhang, and H. Xu, "Evaluation of the integration of retrieval-augmented generation in large language model for breast cancer nursing care responses," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: https://doi.org/10.1038/s41598-024-81052-3
[39] D. D. Prasetya, T. Widiyaningtyas, and T. Hirashima, "Interrelatedness patterns of knowledge representation in extension concept mapping," *Res. Pract. Technol. Enhanc. Learn.*, vol. 20, p. 009, May 2024, doi: https://doi.org/10.58459/rptel.2025.20009