VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i3.6101

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

DIGITALLY FILE EXTRACTION OPTIMISED WITH GPT-4O BASED MOBILE APPLICATION FOR RELEVANT EXERCISE PROBLEM GENERATION

Syanti Irviantina[1*]; Hernawati Gohzaly[2]; Dustin Lionel[3]; Peter Fomas Hia[4]

Faculty of Informatics Engineering[1, 2, 3, 4]
Universitas Mikroskil, Medan, Indonesia[1, 2, 3, 4]
https://mikroskil.ac.id/[1, 2, 3, 4]
syanti@mikroskil.ac.id[1*], hernawati.gohzali@mikroskil.ac.id[2]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *This research studies the creation of an AI-driven question extraction system using the GPT-4o model to improve the accessibility and variety of practice questions for students. The study tackles the difficulties in sourcing relevant practice materials and aims to transform educational technology by integrating mobile learning. A mobile application was built with Dart and Flutter, designed to extract questions from PDF files. The system is capable of generating both multiple-choice and essay questions across different difficulty levels. The quality and relevance of the generated questions were assessed using ROUGE metrics. The results indicated strong performance for multiple-choice questions, especially in single-answer and true/false formats. However, the system encountered difficulties in producing complex essay questions, highlighting the need for further improvements in understanding intricate contextual relationships. Key findings reveal effective generation of multiple-choice questions with high precision and recall; inconsistent performance in essay question generation, with simpler questions yielding better results; and ROUGE-1 metrics surpassing ROUGE-2 and ROUGE-L, indicating a stronger ability to generate straightforward questions. The research concludes that while the developed system shows potential in enhancing educational resources, additional research is necessary to refine complex question generation. Recommendations include broadening the training dataset and creating specialized models for question generation tasks to enhance the effectiveness of AI-assisted learning tools.*

Keywords: *digital file extraction, GPT-4o model, practice question, rouge metrics, question generation.*

Intisari— *Penelitian ini mempelajari pembuatan sistem ekstraksi soal berbasis AI dengan menggunakan model GPT-4o untuk meningkatkan aksesibilitas dan variasi soal latihan bagi siswa. Penelitian ini membahas kesulitan dalam menemukan materi latihan yang relevan dan bertujuan untuk mentransformasi teknologi pendidikan dengan mengintegrasikan pembelajaran mobile. Sebuah aplikasi mobile dibangun dengan Dart dan Flutter, yang dirancang untuk mengekstrak pertanyaan dari file PDF. Sistem ini mampu menghasilkan pertanyaan pilihan ganda dan esai dengan berbagai tingkat kesulitan. Kualitas dan relevansi dari pertanyaan yang dihasilkan dinilai dengan menggunakan metrik ROUGE. Hasilnya menunjukkan kinerja yang baik untuk pertanyaan pilihan ganda, terutama dalam format jawaban tunggal dan benar/salah. Namun, sistem mengalami kesulitan dalam menghasilkan pertanyaan esai yang kompleks, menyoroti perlunya perbaikan lebih lanjut dalam memahami hubungan kontekstual yang rumit. Temuan utama mengungkapkan pembuatan pertanyaan pilihan ganda yang efektif dengan presisi dan recall yang tinggi; kinerja yang tidak konsisten dalam pembuatan pertanyaan esai, dengan pertanyaan yang lebih sederhana memberikan hasil yang lebih baik; dan metrik ROUGE-1 yang melampaui ROUGE-2 dan ROUGE-L, yang mengindikasikan kemampuan yang lebih kuat untuk menghasilkan pertanyaan langsung. Penelitian ini menyimpulkan bahwa meskipun sistem yang dikembangkan menunjukkan potensi dalam meningkatkan sumber daya pendidikan, penelitian tambahan diperlukan untuk menyempurnakan pembuatan pertanyaan yang kompleks.*

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i3.6101

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

Rekomendasi yang diberikan meliputi perluasan dataset pelatihan dan pembuatan model khusus untuk tugas-tugas pembuatan pertanyaan untuk meningkatkan efektivitas alat pembelajaran berbantuan AI.

***Kata Kunci**: ekstraksi file digital, model GPT-4o, soal latihan, metrik rouge, pembuatan pertanyaan.*

INTRODUCTION

In today's evolving society, information technology plays a crucial role in enhancing and reshaping the educational experience for both educators and learners [1]. This technological advancement has transformed the educational landscape by providing broader and more inclusive access to information and diverse learning materials. One significant breakthrough that could revolutionize education is the integration of artificial intelligence (AI) language models like ChatGPT, which has already demonstrated a positive impact on learning outcomes [2].

Despite these advancements, students still face challenges, particularly in accessing a wide range of traditional practice questions. This issue is especially pronounced for those who require additional practice or have varying learning preferences. Students often struggle to find exercises that match their required level of difficulty or relate to the specific material they are studying. This mismatch can hinder the learning experience, dampen motivation, and adversely affect educational effectiveness. Furthermore, a lack of diverse questions can lead to boredom and diminish students' enthusiasm for the subject, ultimately impacting comprehension and retention [3].

Research has shown that learning methods incorporating practice problems are more effective than traditional approaches, such as summarization. Practice problems not only help students apply their knowledge but also enhance critical thinking, problem-solving abilities, and long-term retention of information [4]. However, these benefits can only be fully realized when students have access to a diverse and relevant selection of problems suited to their skill levels.

The use of artificial intelligence-based language models such as ChatGPT has become a significant innovation in the world of education [2]. Research outcomes indicate that artificial intelligence systems such as GPT-4o have the potential to be incorporated into educational environments as analytical tools for narrative examination, potentially enabling both students and instructors to develop a more thorough comprehension of storytelling structures and rhetorical techniques[5]. These characteristics position such AI technology as an ideal candidate for the development of novel and responsive question extraction frameworks. Through the implementation of GPT-4o models in extracting and generating practice questions from digital documentation, there exists significant opportunity to improve both the accessibility and diversity of practice materials available to learners. GPT-4o produces evaluations that demonstrate greater alignment with human assessment criteria, making it a tool with more human-like assessment capabilities [6]. This model can analyze content from various digital sources and generate relevant questions, effectively bridging the gap between learning materials and assessment practices. Additionally, the flexibility of AI technology allows for the customization of question difficulty to meet individual student needs, aligning with the principles of adaptive learning and educational personalization, which are increasingly recognized as essential for enhancing learning effectiveness in the digital age.

ChatGPT's artificial intelligence-based language model has been widely used in education [7], [8], [9], [10], [11]. Research on question development has been conducted in various previous studies, including: the development of exam questions to assess the competence of prospective doctors [12], the creation of practice questions for programming learning [13], and the development of various questions that are semantically similar but lexically different to assess the same concept [14]. The implementation of m-learning positively enhances learning flexibility, enabling access to materials anytime and anywhere. Teachers and students acknowledge significant benefits in improving learning outcomes and material comprehension [15] . A lot of research has been conducted in relation to the use of mobile learning applications to do practice questions [16], [17], [18].

This research aims to investigate the potential of utilizing the GPT-4o model to develop a question extraction system from digital files. The focus of this study will be on how this system can effectively address accessibility challenges, offer a diverse array of practice questions, and enhance both learning performance and the assessment of student knowledge. To evaluate system performance, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) method is implemented as a measurement standard. ROUGE is an evaluation technique widely applied to assess

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

and compare the effectiveness of automatic summarization systems. This method functions to evaluate the degree of correspondence between system-generated summaries and the information contained in the original documents. ROUGE scores are calculated by comparing the summary output produced by a system or specific method with summaries manually created by humans (human annotators), thus obtaining an objective assessment of the system's performance [19], [20], [21]. , and has been instrumental in comparing the performance of prominent large language models such as GPT [22].

MATERIALS AND METHODS

This research employs an application development approach that incorporates the GPT-4o model. The primary aim is to enhance the text extraction process and assess the quality and relevance of the generated questions. For text extraction, this study relies on digital text files in PDF format.

Problem Identification

The research addresses the limitations in accessibility and variety of practice questions in the student learning process, as well as the difficulty in obtaining relevant practice questions that align with the studied material and desired difficulty level.

Data Collection

A systematic literature review is conducted, focusing on natural language processing and GPT models. The study explores ROUGE metrics and their application in evaluating generated text. Additionally, mobile application development techniques are investigated. Interviews with students and teachers are conducted to gather information on question creation techniques in educational technology. The dataset consists of four digital PDF files of varying sizes and page lengths: 2.01 MB (51 pages), 10 MB (321 pages), 35.3 MB (296 pages), and 36 MB (216 pages). These documents were specifically selected to represent diverse academic content across different file sizes and lengths, ensuring a comprehensive analysis of the text extraction and processing capabilities of the proposed system.

GPT-4o based Mobile Application Development

The development process involves several critical steps:

1. Selecting an appropriate mobile platform
2. Designing an interface to provide an optimal user experience
3. Creating a prototype to visualize and test application features
4. Conducting prototype testing to gather feedback and identify areas for improvement prior to final implementation

GPT-4o Model Integration

The integration process encompasses the implementation of the GPT-4o model into the mobile application, as well as the development of prompts and strategies to generate effective questions. This research leverages GPT-4o's prompt engineering capabilities to ensure optimal question generation results. Through meticulously designed prompts, the system can produce questions that are both contextually relevant and pedagogically sound. Several key reasons for utilizing prompt engineering in this system include: its ability to control output parameters in detail, consistency in the quality of generated questions, capability to adjust across various difficulty levels, and enhanced efficiency in learning content development processes.

Application Testing

The testing process incorporates quantitative evaluation using ROUGE Metrics [16] to analyze the quality of text extraction and generated questions, as well as their relevance to the content of the extracted digital files. ROUGE Metrics, commonly used in Natural Language Generation (NLG), measure the correspondence between machine-generated text and human-written reference text. The metric functions by calculating different types of scores - ROUGE-1, ROUGE-2, and ROUGE-L - with higher scores indicating superior performance, thus providing a straightforward methodology for comparing generated outputs [23]. The metrics used in this testing include:

1. Recall: Calculates the number of overlapping n-grams between the model output and reference text, divided by the total n-grams in the reference [19].
2. Precision: Computes the number of matching n-grams divided by the total n-grams produced by the model [19].
3. F1 Score: Combines precision and recall into a single metric by calculating their harmonic mean [19].

Additionally, the application underwent testing with 35 high school students, focusing on two key aspects:

VOL. 10. NO. 3 FEBRUARY 2025
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i3.6101

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

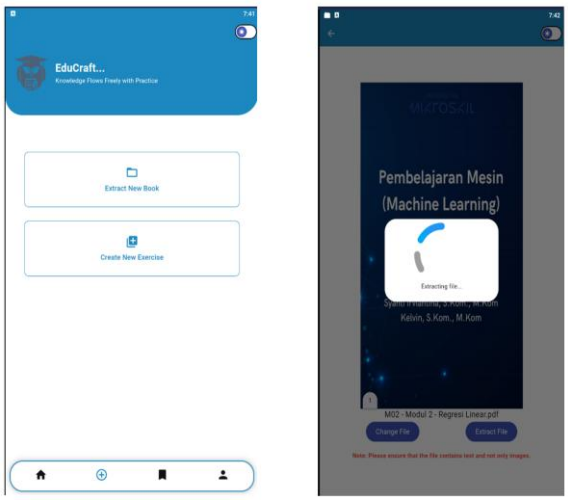
1. File upload and extraction process: Students tested the efficiency and accuracy of the file upload mechanism and the text extraction functionality by uploading various document formats and evaluating the extraction results.
2. Question generation capability: The application's ability to generate practice questions with varying levels of complexity was assessed based on student responses to the generated questions and their perception of difficulty alignment.
3. Learning experience perspective: Students evaluated the application's contribution to their overall learning process, focusing on how well it supported knowledge acquisition and retention. This assessment examined the quality of answer explanations, conceptual understanding support, feedback mechanisms, and progress tracking features to determine if the application functioned merely as a question generator or as a comprehensive learning tool.

Conclusion

The research concludes by drawing inferences from the testing process, presenting the final results of the study.

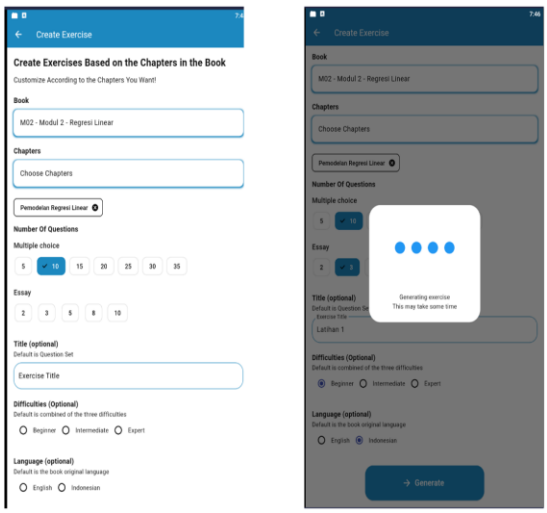
RESULTS AND DISCUSSION

In developing this digital file extraction application, we used the Dart programming language and the Flutter open-source framework. At the initial stage, the application is designed to accept input in the form of digital files in pdf formats, which will then be extracted by the implemented model, as shown in Figure 1.



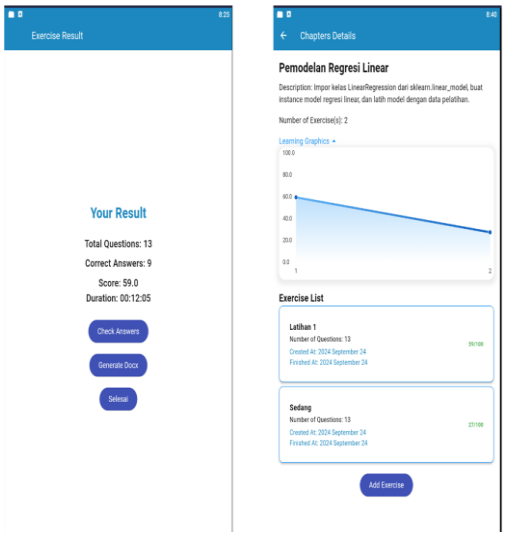
Source: (Research Results, 2024)
Figure 1. File Extraction Mobile Application

The results of this application processing will produce various levels of questions, ranging from easy, medium, to difficult, as well as two types of practice questions, namely multiple choice and essay. In addition, users can determine the number of questions as desired, or let the system determine it automatically and can choose the language, namely English or Indonesian as shown in Figure 2 below.



Source: (Research Results, 2024)
Figure 2. Determine Number and Language

Users can work on the practice questions and receive correction results if there are wrong answers after all the questions are done and submitted. The results of working on practice questions will be presented in the form of visualisations to monitor the progress of students' abilities. This can be seen in Figure 3 below.



Source: (Research Results, 2024)
Figure 3. Exercise Question Extraction Results

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

The testing process involved several stages, including analysing the quality of the questions generated, assessing the relevance to the context of the extracted data, and measuring the response time of the system in generating questions based on user input as shown in Table 1.

Table 1. Digital File Test List

File	File Size	Topic / Chapter	File Extraction Time
A	2.01 MB	Mandatory Attributes and Identity Attributes	27000 ms
B	10 MB	Chapter 6 - Presenting and Publishing Research	32990 ms
C	35.3 MB	Bab I and Bab II	44833 ms
D	36 MB	Bab 3 - Pressure	42175 ms

Source: (Research Results, 2024)

Testing was conducted on each digital file for multiple-choice question types, which were divided into three types:

1. Single answer choice: questions with one correct answer.
2. Multiple-choice questions: questions with more than one correct answer.
3. True/false questions: questions that require a response in the form of a correct or incorrect answer.

This research aims to analyse the quality of the generated queries and asses their relevance to the context of the extracted data, using the Rouge Metric method. This assessment includes precision (P), Recall (R) dan F1-Score (F) measurements, as shown in Table 2,3,4 below:

Table 2. Results for Multiple Choice Questions with Rouge-1

File	Multiple choice	Rouge-1	File	Multiple choice	Rouge-1
A	Single answer choice	P: 0.88 R: 0.66 F: 0.76	C	Single answer choice	P: 0.62 R: 0.52 F: 0.57
	Plural answer choices	P: 0.66 R: 0.93 F: 0.77		Plural answer choices	P: 0.93 R: 0.53 F: 0.68
	True/False	P: 0.70 R: 0.63 F: 0.66		True/False	P: 0.64 R: 0.42 F: 0.51
B	Single answer choice	P: 0.73 R: 0.64 F: 0.68	D	Single answer choice	P: 0.73 R: 0.64 F: 0.68
	Plural answer choices	P: 0.70 R: 0.58 F: 0.63		Plural answer choices	P: 0.66 R: 0.57 F: 0.61
	True/False	P: 0.91 R: 0.84 F: 0.88		True/False	P: 0.87 R: 1.00 F: 0.93

Source: (Research Results, 2024)

Table 3. Results for Multiple Choice Questions with Rouge-2

File	Multiple choice	Rouge-2	File	Multiple choice	Rouge-2
A	Single answer choice	P: 0.37 R: 0.27 F: 0.31	C	Single answer choice	P: 0.40 R: 0.33 F: 0.36
	Plural answer choices	P: 0.60 R: 0.85 F: 0.70		Plural answer choices	P: 0.80 R: 0.44 F: 0.57
	True/False	P: 0.33 R: 0.30 F: 0.31		True/False	P: 0.23 R: 0.15 F: 0.18
B	Single answer choice	P: 0.28 R: 0.25 F: 0.26	D	Single answer choice	P: 0.42 R: 0.37 F: 0.40
	Plural answer choices	P: 0.42 R: 0.34 F: 0.38		Plural answer choices	P: 0.43 R: 0.37 F: 0.40
	True/False	P: 0.27 R: 0.25 F: 0.26		True/False	P: 0.85 R: 1.00 F: 0.92

Source: (Research Results, 2024)

Table 4. Results for Multiple Choice Questions with Rouge-L

File	Multiple choice	Rouge-L	File	Multiple choice	Rouge-L
A	Single answer choice	P: 0.77 R: 0.58 F: 0.66	C	Single answer choice	P: 0.50 R: 0.42 F: 0.45
	Plural answer choices	P: 0.66 R: 0.93 F: 0.77		Plural answer choices	P: 0.93 R: 0.53 F: 0.68
	True/False	P: 0.70 R: 0.63 F: 0.66		True/False	P: 0.64 R: 0.42 F: 0.51
B	Single answer choice	P: 0.66 R: 0.58 F: 0.62	D	Single answer choice	P: 0.66 R: 0.58 F: 0.62
	Plural answer choices	P: 0.45 R: 0.37 F: 0.40		Plural answer choices	P: 0.63 R: 0.55 F: 0.59
	True/False	P: 0.58 R: 0.53 F: 0.56		True/False	P: 0.87 R: 1.00 F: 0.93

Source: (Research Results, 2024)

Evaluation results using the Rouge metric show that the model performs well in generating multiple-choice questions, especially for multiple-choice and true/false question types. The data processing presented in the table includes precision and recall values for various question categories.

The Rouge-1 metric showed positive results, with 88% precision and 66% recall for multiple-choice questions, indicating that the model was able to generate highly relevant and accurate questions in this category. For true/false questions, the model recorded a precision of 70% and a recall of 63%, which also indicates a good performance in generating questions that are appropriate and in context.

However, there was a significant drop in the Rouge-2 metric, where precision and recall were lower compared to the Rouge-1 metric. This suggests that the model faces difficulties in capturing more complex bigram relationships, possibly due to a lack of data or variations in the structure of the generated questions.

The next test was conducted on each digital file for the essay question type, where each question consisted of two essay questions. The testing process starts with the researcher (P) manually creating the questions and then manually searching for answers from the reference source (R). The results obtained by the model (HM) are then compared. The analysis of the quality of the generated questions and the assessment of relevance to the context of the extracted data was carried out using the Rouge Metric method, as seen in Table 5, 6, 7 below:

Table 5. Test results of Essay Questions with Rouge-1

File	Question 1	Rouge-1	Question 2	Rouge-1
A	P R HM	Precision: 0.78, Recall: 0.56, F-measure: 0.65	P R HM	Precision: 0.67, Recall: 0.87, F-measure: 0.75
B	P R HM	Precision: 0.72, Recall: 0.37, F-measure: 0.49	P R HM	Precision: 0.56, Recall: 0.29, F-measure: 0.38
C	P R HM	Precision: 0.65, Recall: 0.71, F-measure: 0.68	P R HM	Precision: 0.89, Recall: 0.51, F-measure: 0.65
D	P R HM	Precision: 0.68, Recall: 0.42, F-measure: 0.52	P R HM	Precision: 0.62, Recall: 0.43, F-measure: 0.51

Source: (Research Results, 2024)

Table 6.Test results of Essay Questions with Rouge-2

File	Question 1	Rouge-2	Question 2	Rouge-2
A	P R HM	Precision: 0.65, Recall: 0.46, F-measure: 0.54	P R HM	Precision: 0.52, Recall: 0.68, F-measure: 0.59
B	P R HM	Precision: 0.33, Recall: 0.17, F-measure: 0.22	P R HM	Precision: 0.23, Recall: 0.12, F-measure: 0.15
C	P R HM	Precision: 0.39, Recall: 0.43,	P R HM	Precision: 0.64, Recall: 0.36,
D	P R HM	F-measure: 0.41 Precision: 0.26, Recall: 0.16, F-measure: 0.20	P R HM	F-measure: 0.46 Precision: 0.40, Recall: 0.28, F-measure: 0.33

Source: (Research Results, 2024)

Table 7. Test Results of Essay Questions with Rouge-L

File	Question 1	Rouge-L	Question 2	Rouge-L
A	P R HM	Precision: 0.78, Recall: 0.56, F-measure: 0.65	P R HM	Precision: 0.67, Recall: 0.87, F-measure: 0.75
B	P R HM	Precision: 0.53, Recall: 0.27, F-measure: 0.36	P R HM	Precision: 0.39, Recall: 0.20, F-measure: 0.27
C	P R HM	Precision: 0.54, Recall: 0.60, F-measure: 0.57	P R HM	Precision: 0.86, Recall: 0.49, F-measure: 0.63
D	P R HM	Precision: 0.50, Recall: 0.31, F-measure: 0.38	P R HM	Precision: 0.54, Recall: 0.38, F-measure: 0.44

Source: (Research Results, 2024)

In general, the model showed significant variation in performance in generating essay questions. The precision, recall and F-measure metrics give an idea of how well the model does in generating relevant and accurate questions.
Some questions, such as the first and second questions, showed relatively good results, indicating that the model was able to generate relevant questions. However, in the fourth question, the model showed low performance on the Rouge-1 metric, indicating difficulty in generating relevant questions.

Meanwhile, the seventh and eighth questions also recorded low scores on the Rouge-2 metric, reflecting the model's difficulty in capturing complex bigram relationships. Overall, the Rouge-1 metric gave reasonably good results for most questions, while the Rouge-2 and Rouge-L metrics showed significant drops in precision and recall, especially for the more difficult questions. This suggests that the model may be more effective in generating simple questions compared to more complex questions.

Based on the Rouge metric testing results presented in the research, several factors can be

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

identified that explain why the model experiences difficulties in generating complex essay questions.

Performance decline in bigram metrics (Rouge-2)

Analysis of Rouge-2 metric results shows a significant decrease compared to Rouge-1 metrics, especially for essay questions. For example, in File B question 2, the F-measure value drops from 0.38 (Rouge-1) to 0.15 (Rouge-2). This dramatic decrease indicates that:

1. Difficulty in maintaining linguistic structure: the model tends to preserve individual words (unigrams) from the source text but struggles to maintain relationships between words (bigrams) that form complex syntactic structures.
2. Limitations in understanding semantic context: the low Rouge-2 scores indicate the model has difficulty capturing deeper semantic relationships between phrases in the source text.

Correlation analysis between file size and content complexity

There is an interesting pattern when comparing model performance based on file size and complexity:

1. File A (2.01 MB) shows the highest performance on essay questions (Rouge-1 F-measure: 0.65-0.75)
2. File B (10 MB) shows the lowest performance (Rouge-1 F-measure: 0.38-0.49)
3. Files C and D (>35 MB) show intermediate performance

This demonstrates that model performance is not solely influenced by file size, but more significantly affected by:

1. Structural complexity of text : files with more complex linguistic structures produce lower scores
2. Information density : files with high information density ratios (many concepts in relatively little text) make it difficult for the model to generate comprehensive questions

Analysis of long-range sequential modeling limitations

The significant difference between Rouge-L and Rouge-1 or Rouge-2 metrics provides important insights. In File B question 1, the Rouge-L F-measure (0.36) is much lower than Rouge-1 (0.49), indicating:

1. Difficulty in maintaining long-term coherence: the model struggles to maintain logical flow in generating essay questions that require understanding of extended argumentative structures
2. Limitations in capturing long-range dependencies : the model demonstrates limitations in capturing distant dependency relationships that are crucial for complex essay question

These findings highlight the importance of developing domain-specific fine-tuning techniques and more sophisticated discursive structure processing mechanisms to enhance the model's ability to generate high-quality complex essay questions.

Further testing with 35 high school students revealed promising results regarding usability and effectiveness. The file upload and extraction process was notably efficient, with 57.1% of students completing this process within 1-3 minutes. Approximately 68.6% of students appreciated the application's capacity to generate questions with varying complexity levels.

Regarding question diversity, 54.3% of respondents rated this aspect as excellent, noting the inclusion of both multiple-choice and essay questions. Content relevance was also highly rated, with 65.7% of students stating that the generated questions aligned well with their course materials. Additionally, 71.4% of students found the difficulty level of questions to be appropriate for their abilities, being neither too easy nor too challenging.

From a learning experience perspective, 60% of students rated the answer explanations provided as very helpful in understanding concepts. 65.7% of respondents felt that the application not only generated questions but also supported their process of understanding the material. The automated grading system was considered clear by 54.3% of students, while 51.4% found it easy to track their learning progress through the application.

CONCLUSION

This research has demonstrated the development and implementation of a question generation system based on the GPT-4o model, offering significant practical applications in educational contexts. The developed mobile application successfully generates multiple question types (single-choice, multiple-choice, true/false, and essay questions) from PDF documents, providing an accessible tool for both educators and students. Evaluation using ROUGE metrics revealed that the system performs well for structured questions, particularly multiple-choice

and true/false formats, with high precision and recall values indicating strong alignment with reference questions.

However, performance analysis highlighted important limitations in generating complex question formats. The system showed notable degradation in ROUGE-2 and ROUGE-L metrics when tasked with producing essay questions that require deeper conceptual understanding and more sophisticated linguistic structures. This indicates that while the model effectively captures unigram relationships (ROUGE-1), it struggles with more complex bigram relationships and maintaining extended sequence relevance essential for higher-order questioning.

The practical implications of this technology are substantial for educational stakeholders. For teachers, the system offers significant efficiency gains, reducing assessment preparation time while maintaining quality and enabling differentiated instruction through varied question complexity. Students benefit through enhanced self-directed learning capabilities, allowing them to generate personalized practice materials from their own study resources and identify knowledge gaps prior to formal assessments.

Future enhancement pathways include both data-driven and architectural approaches. Enlarging and diversifying the training dataset represents an immediate strategy to improve question variety and relevance.

This research contributes to the evolving field of AI-assisted education while acknowledging current technical limitations. The findings suggest that while language model-based question generation systems show immediate practical utility, continued research is needed to develop specialized models for educational question generation that better capture complex cognitive relationships and disciplinary nuances.

REFERENCE

[1] F. M. Sinaga, S. Irviantina, and S. J. Pipin, "Pelatihan Pembuatan Konten Pembelajaran Berbasis Video pada SMA Methodist 6," *Journal of Social Responsibility Projects by Higher Education Forum*, vol. 4, no. 3, pp. 139–144, Mar. 2024, doi: 10.47065/jrespro.v4i3.4588.

[2] M. Montenegro-Rueda, J. Fernández-Cerero, J. M. Fernández-Batanero, and E. López-Meneses, "Impact of the Implementation of ChatGPT in Education: A Systematic Review," *Computers*, vol. 12, no. 8, p. 153, Jul. 2023, doi: 10.3390/computers12080153.

[3] P. Reilly, Z. Al Khuridah, R. Cooper, L. L. Hsu, and C. Hentea, "Commercial Question Banks Present a Limited Scope of Sickle Cell Disease," *Blood*, vol. 140, no. Supplement 1, pp. 10779–10780, Nov. 2022, doi: 10.1182/blood-2022-162751.

[4] M. L. Rivers, "Test Experience, Direct Instruction, and Their Combination Promote Accurate Beliefs about the Testing Effect," *Journal of Intelligence*, vol. 11, no. 7, p. 147, Jul. 2023, doi: 10.3390/jintelligence11070147.

[5] H. Geng and H. Wei, "Exploring ChatGPT's Capabilities in Creative Writing: Can GPT-4o Conduct Rhetorical Move Analysis in Narrative Short Stories?," *ASEAN Journal of Applied Linguistics*, vol. 3, no. 1, pp. 44–59, 2024.

[6] S. Donthi *et al.*, "Improving LLM Abilities in Idiomatic Translation," Jul. 2024, doi: 10.48550/arXiv.2407.03518.

[7] S.-V. Fulgencio, "Developing Effective Educational Chatbots with GPT: Insights from a Pilot Study in a University Subject," *Trends in Higher Education*, vol. 3, no. 1, pp. 155–168, Mar. 2024, doi: 10.3390/higheredu3010009.

[8] J. Savelka, A. Agarwal, C. Bogart, Y. Song, and M. Sakr, "Can Generative Pre-trained Transformers (GPT) Pass Assessments in Higher Education Programming Courses?," *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, pp. 117–123, Jun. 2023, doi: 10.1145/3587102.3588792.

[9] W. Suharmawan, "Pemanfaatan Chat GPT Dalam Dunia Pendidikan," *Education Journal : Journal Educational Research and Development*, vol. 7, no. 2, pp. 158–166, 2023, doi: 10.31537/ej.v7i2.1248.

[10] M. Mustafa, "Aktivitas Siswa dalam Memecahkan Masalah Matematika dengan Berpikir Komputasi Berbantuan Chat-GPT," *MATHEMA: JURNAL PENDIDIKAN MATEMATIKA*, vol. 5, no. 2, pp. 283–298, Oct. 2023, doi: 10.33365/jm.v5i2.3469.

[11] A. Setiawan and U. K. Luthiyani, "Penggunaan ChatGPT Untuk Pendidikan di Era Education 4.0: Usulan Inovasi Meningkatkan Keterampilan Menulis ," *Jurnal PETISI (Pendidikan Teknologi Informasi)*, vol. 04, no. 01, pp. 49–58, Feb. 2023.

[12] S. Bedi et al., "QUEST-AI: A System for Question Generation, Verification, and Refinement using AI for USMLE-Style

Exams," *Biocomputing 2025*, pp. 54–69, Nov. 2024, doi: 10.1142/9789819807024_0005.

[13] I. Hsiao and C.-Y. Chung, "AI-Infused Semantic Model to Enrich and Expand Programming Question Generation," *Journal of Artificial Intelligence and Technology*, vol. 2, no. 2, Mar. 2022, doi: 10.37965/jait.2022.0090.

[14] M. Rathod, T. Tu, and K. Stasaski, "Educational Multi-Question Generation for Reading Comprehension," *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, doi: 10.18653/v1/2022.bea-1.26.

[15] B. Ananda and S. Suranto, "Transformasi Pembelajaran di Sekolah Menengah Kejuruan: Analisis Mendalam Fleksibilitas M-learning," *Ideguru: Jurnal Karya Ilmiah Guru*, vol. 9, no. 2, pp. 695–701, Jan. 2024, doi: 10.51169/ideguru.v9i2.936.

[16] B. Jerome, R. Van Campenhout, and B. G. Johnson, "Automatic Question Generation and the SmartStart Application," *Proceedings of the Eighth ACM Conference on Learning @ Scale*, pp. 365–366, Jun. 2021, doi: 10.1145/3430895.3460878.

[17] E. Kristanti, G. I. Kharisma, and N. P. Sari, "Pelatihan Penyusunan Soal Berbasis Mobile Learning Sebagai Upaya Menghadapi Era Pendidikan 4.0," *WIDYA LAKSANA*, vol. 10, no. 1, pp. 59–65, Mar. 2021, doi: 10.23887/jwl.v10i1.28915.

[18] A. Widyatama and F. W. Pratama, "Pengembangan Mobile Learning PINTHIR Berbasis Android sebagai Sumber Belajar dan Sarana Mengerjakan Soal Trigonometri SMA," *Mosharafa: Jurnal Pendidikan Matematika*, vol. 11, no. 1, pp. 25–36, Jan. 2022, doi: 10.31980/mosharafa.v11i1.684.

[19] Halimah, S. Agustian, and S. Ramadhani, "Peringkasan Teks Otomatis (Automated Text Summarization) Pada Artikel Berbahasa indonesia Menggunakan Algoritma Lexrank," *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 3, no. 3, pp. 371–381, Dec. 2022.

[20] A. Yogi Setiawan, I. G. Mahendra Darmawiguna, and G. Aditra Pradnyana, "Sentiment Summarization Evaluasi Pembelajaran Menggunakan Algoritma LSTM (Long Short Term Memory)," *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)* , vol. 11, no. 2, pp. 183–191, Aug. 2022.

[21] F. Noprianto, S. Agustian, and M. Irsyad, "Clustering Peringkasan Teks Otomatis Dokumen Berita Menggunakan Metode K-Means," in *Sendiko - Prosiding Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat Bidang Ilmu Komputer*, R. Pamungkas, Saifulloh, and Andria, Eds., Madiun: Universitas PGRI Madiun, Jun. 2023, pp. 139–147.

[22] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," *SSRN Electronic Journal*, May 2022, doi: 10.2139/ssrn.4120317.

[23] R. Harang, "Beyond ROUGE: Applying an ELO algorithm to rank model performances in summarization," In *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing*, pp. 2799-2804, 2024.