JEDUMATIC

e-ISSN 3031-8408

Journal of Education and Informatics Research /Vol.7, No.1 2026

How Well Do LLMs Understand Indonesian Textbooks? An Evaluation on Informatics Question Answering

Nur Aisyah

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
email: 6025241004@student.its.ac.id

ABSTRAK

Kemunculan *Large Language Models* (LLM) membuka peluang besar dalam pengembangan sistem tanya jawab otomatis di bidang pendidikan. Penelitian ini membandingkan kualitas generasi teks antara *GPT-4o-mini* dan *DeepSeek-v4-flash* pada tugas *question answering* (QA) berbasis konteks menggunakan buku teks Informatika untuk SMA/MA kelas X, XI, dan XII yang bersumber dari Sistem Perbukuan Indonesia (SISPBK). Dataset yang dibangun terdiri dari 1.760 pasang konteks-pertanyaan-jawaban yang mencakup enam area topik sesuai Kurikulum Merdeka. Kedua model dievaluasi menggunakan *zero-shot prompting* tanpa *fine-tuning*. Evaluasi dilakukan secara otomatis menggunakan *LLM-as-a-Judge* dengan *DeepSeek* sebagai *evaluator*, yang menilai tiga aspek kualitas: *answerability* (kesesuaian dengan konteks), *fluency* (kualitas kebahasaan), dan *relevance* (keterkaitan dengan pertanyaan) menggunakan skala Likert 1-5. Hasil menunjukkan bahwa kedua model mencapai skor *fluency* tertinggi dibandingkan aspek lainnya, di mana *GPT-4o-mini* memperoleh skor 4,60-5,00 dan *DeepSeek-v4-flash* memperoleh skor 4,00-5,00 di seluruh jenjang kelas, mengindikasikan kemampuan yang baik dalam menghasilkan teks Bahasa Indonesia yang gramatikal dan alami. *GPT-4o-mini* unggul dalam aspek *answerability*, sedangkan *DeepSeek-v4-flash* lebih baik dalam aspek *relevance*. Namun, skor *answerability* kedua model yang konsisten di bawah 3,00 menunjukkan keterbatasan dalam memenuhi persyaratan QA *closed-book* berbasis kurikulum. Temuan ini mengindikasikan perlunya penyempurnaan desain *prompt* dan *fine-tuning* berbasis domain untuk meningkatkan efektivitas LLM dalam aplikasi pendidikan di Indonesia.

Kata Kunci: Informatika, Kurikulum Merdeka, *Large Language Model*, *LLM-as-a-Judge*, Pertanyaan-Jawaban

ABSTRACT

The emergence of Large Language Models (LLMs) presents significant opportunities for developing automatic question answering (QA) systems in education. This study compares the text generation quality of GPT-4o-mini and DeepSeek-v4-flash on context-grounded QA tasks using official Informatics textbooks for Senior High School (SMA/MA) grades X, XI, and XII sourced from Indonesia's National Textbook System (SISPBK). A dataset of 1,760 context-question-answer triples was constructed, covering six major topic areas aligned with the Merdeka Curriculum. Both models were evaluated under identical zero-shot prompting conditions without any task-specific fine-tuning. Automated evaluation was conducted using an LLM-as-a-Judge framework with DeepSeek as the evaluator, assessing three quality dimensions: answerability (contextual grounding), fluency (linguistic quality), and relevance (topical alignment) on a 1-5 Likert scale. Results show that both models achieved the highest scores in the fluency dimension, with GPT-4o-mini scoring between 4.60 and 5.00 and DeepSeek-v4-flash scoring between 4.00 and 5.00 across all grade levels, demonstrating strong capability in producing grammatically correct and natural Bahasa Indonesia. GPT-4o-mini outperformed in answerability, while DeepSeek-v4-flash showed a slight advantage in relevance. However, consistently low answerability scores below 3.00 for both models reveal limitations in meeting the strict requirements of closed-book, curriculum-based QA settings. These findings highlight the need for improved prompt design and domain-specific fine-tuning to enhance LLM effectiveness in Indonesian secondary education.

Keywords: Informatics, Large Language Model, LLM-as-a-Judge, Merdeka Curriculum, Question Answering

INTRODUCTION

The emergence of Large Language Models (LLMs) has become one of the most significant developments in artificial intelligence over the past decade. Based on the Transformer architecture introduced by Ashish Vaswani, et al. (Vaswani et al., 2017), LLMs have evolved from simple language models into advanced systems capable of understanding, reasoning, and generating natural language in ways that closely resemble human communication. This architectural advancement, combined with training on large-scale internet data and alignment techniques, has equipped modern LLMs to effectively perform a wide variety of tasks has equipped modern LLMs to effectively perform a wide variety of tasks, including text summarization, translation, code generation, mathematical reasoning, and open-domain question answering (Yang et al., 2024). Their growing capabilities and flexibility have encouraged adoption across many important sectors, with education emerging as one of the most challenging areas due to the strong potential of LLMs to support teaching and learning processes.

In the educational domain, LLMs provide significant opportunities for the development of automatic question-answering (QA) systems. Traditional QA systems mainly depended on information retrieval methods and rule-based answer extraction, which limited their flexibility and ability to adapt to different learning contexts (Upadhyay et al., 2024). In contrast, LLMs can generate fluent, context-aware, and educationally appropriate responses without requiring extensive manual rule design. This capability is particularly important for curriculum-aligned QA systems, where responses must be not only factually accurate but also consistent with the vocabulary, learning scope, and conceptual depth defined in official educational materials. As a result, the integration of LLMs into educational technologies, such as intelligent tutoring systems, automated formative assessment platforms, and digital learning assistants, has attracted growing attention from both researchers and industry worldwide (Yan et al., 2024).

Two of the most widely recognized LLMs available today are OpenAI's (GPT-4o-mini) and DeepSeek AI's (DeepSeek-v4-flash). GPT-4o-mini is the latest model in OpenAI's generative pre-trained transformer series and features a multimodal architecture that can process text, images, and audio within a single system (Upadhyay et al., 2024). In contrast, DeepSeek-v4-flash adopts a Mixture-of-Experts (MoE) architecture with 236 billion total parameters, while activating only 21 billion parameters during each forward pass, allowing the model to balance high performance with computational efficiency (Yan et al., 2024). Both models have achieved strong results on standard natural language processing (NLP) benchmarks and support multilingual text generation, including Bahasa Indonesia, making them promising candidates for educational applications in Indonesia. However, their performance and response characteristics on domain-specific, curriculum-aligned educational content in Indonesian language have not yet been thoroughly investigated.

In Indonesia, the adoption of AI-powered educational tools is closely connected to the country's broader digital transformation. The Ministry of Education, Culture, Research, and Technology has introduced various initiatives to modernize the national education system, including the implementation of the Merdeka Curriculum, which emphasizes project-based learning, student-centered education, and the development of twenty-first century skills (Indarta et al., 2022). As part of this initiative, Kemendikbudristek has provided free access to official digital textbooks through the National Textbook System (Sistem Perbukuan Indonesia/SISPBK), a centralized platform containing curriculum-aligned learning materials for various educational levels and subjects (Kemendikbristek, 2024). These textbooks have undergone comprehensive content review processes and are regarded as reliable and authoritative learning resources for secondary education in Indonesia.

Informatics holds strategic importance within the subjects available in SISPBK. Informatics textbooks for Senior High School under the Merdeka Curriculum cover a wide range of topics, including

computational thinking, algorithms and programming, computer networks, database systems, information security, and basic artificial intelligence. These topics are closely related to the skills needed in an increasingly digital economy and support Indonesia's goal of developing a digitally literate and technologically skilled workforce. In addition, the textbooks are organized in a clear chapter-based structure and are publicly accessible, making them a suitable resource for developing domain-specific question-answering (QA) datasets and evaluating LLM-generated responses using curriculum-aligned educational content.

Evaluating LLM-generated text in educational settings presents challenges that differ from general natural language processing (NLP) benchmarking. Traditional automatic evaluation metrics such as BLEU (Wentzel, 1922) and ROUGE (Sollers, 2014) rely on lexical overlaps and have been shown to correlate weakly with human judgments of answer quality, especially in open-ended generative tasks. Although human evaluation is generally more reliable, it is often too expensive and time-consuming to be applied at the scale required for comprehensive model comparisons. This limitation has led to the development of the LLM-as-a-Judge approach, where a capable LLM is used as an automated evaluator to assess and compare model outputs across multiple quality dimensions (Chang et al., 2024). When properly prompted, LLM-based evaluators have demonstrated strong agreement with human judgments and have been effectively used in various evaluation settings, including conversational response quality and domain-specific factual accuracy.

Despite these converging developments, a significant research gap remains. To the best of our knowledge, no prior study has systematically compared GPT-4o-mini and DeepSeek-v4-flash on context-grounded question-answering (QA) generation using Indonesian-language educational content derived from official national curriculum materials. Existing comparative studies of LLMs in educational contexts have largely focused on English-language datasets (Kwak & Pardos, 2024), while research in Indonesian natural language processing (NLP), although growing, has not yet produced benchmarks specifically designed for secondary school Informatics or for evaluating alignment with curriculum documents. Moreover, the application of the LLM-as-a-Judge framework for assessing Indonesian educational QA responses across multiple evaluation dimensions simultaneously remains largely unexplored in the literature.

This study is designed to comprehensively address these gaps. Our main contributions are twofold: first, we construct a new publicly releasable question-answering (QA) dataset consisting of 1,760 context-question-answer triples generated from official Informatics textbooks for Senior High School Grade X, XI and XII sourced from SISPBK, covering six major topic areas aligned with the Merdeka Curriculum; second, we conduct a controlled comparative evaluation of GPT-4o-mini and DeepSeek-v2-flash on the task of context-grounded question generation and answering under identical zero-shot prompting conditions. The evaluation is further supported by a three-dimensional LLM-as-a-Judge framework using DeepSeek as the evaluator, assessing responses based on answerability (contextual grounding), fluency (linguistic quality), and relevance (topical alignment). The results are expected to provide practical guidance for deploying LLM-based QA systems in Indonesian secondary education and to serve as a reusable benchmark for future comparative studies.

METHOD

This section explains the research methodology used in this study. The methodology consists of three main stages: (1) collecting and building the dataset from official Informatics textbooks obtained from SISPBK, (2) designing the experiment, including the prompting and text generation settings for both models, and (3) conducting automated evaluation using an LLM-as-a-Judge framework. Fig 1. shows the overall research workflow, and each stage is described in detail in the following subsections.

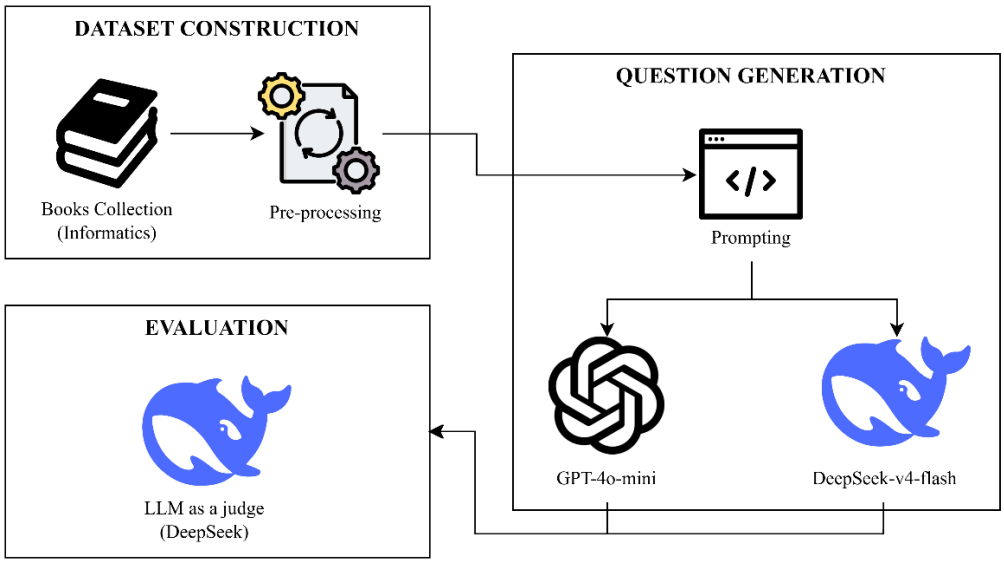


Fig 1. Research workflow on this proposed method

1. Dataset

The main data source for this study is a selected collection of official Informatics textbooks provided by the Indonesian government through the SISPBK portal (https://buku.kemdikbud.go.id), managed by the Ministry of Education, Culture, Research, and Technology. The textbooks used in this study cover Informatics subjects for Senior High School at grades X, XI, and XII, resulting in a total of six textbook volumes, with three volumes for each educational track. All textbooks are published under an open government license, making them freely accessible to the public, and have undergone formal review processes to ensure alignment with the national Merdeka Curriculum. These textbooks were selected because they serve as the main learning reference for Informatics education at the secondary school level in Indonesia and cover a wide range of topics (see Table 1).

Table 1. Topics mapping across grades in informatics textbooks

Grade	Topics
X	Data, Information, and Validation, Algorithms and Data Structures, Technology and Digital Culture, Information Security and Management
XI	Digital Information, Data Analysis, Algorithmic Strategies, Data Structure Design, and Solution Analysis, Developing Computational Thinking Capabilities and Algorithm Implementation, Designing Computer Networks and Their Security Configurations
XII	Social Impact of Informatics, Software Engineering, Database Systems, Artificial Intelligence

Before the textbook content could be used for question generation, a content extraction process was carried out to convert the source documents from their original digital formats into plain text. This step is important because LLM-based generation systems require raw text input rather than formatted document files. In this study, the extraction process was implemented programmatically to handle the PDF format. The files, the system extracted readable text page by page in sequence. Automated text extraction from government-issued textbooks can be challenging because these documents often contain complex layouts such as multi-column formatting, tables, figures with captions, sidebar notes, and decorative elements. These components may reduce the accuracy of plain-text extraction. As a result, the extracted text was not always perfectly clean, and some noise, such as fragmented sentences near figures or improperly aligned table content, was expected. These issues were addressed in the following processing stages. The result of this process is a knowledge base that includes three components (triplets)

namely context, question and cognitive level. Context is a collection of material from several sub-themes in the textbook. Questions a collection of questions created based on context. Cognitive level is a label of cognitive based on the generated questions.

2. Question Generation

This study employs a controlled experimental design to compare the text generation performance of GPT-4o-mini and DeepSeek-v4-flash using informatics textbooks. Both models are evaluated under the same zero-shot and context-based conditions. In this setting, each model is required to generate answers only from the provided textbook passage, without using prior examples or external knowledge. To maintain fairness and avoid bias, neither model receive fine-tuning nor task-specific adaptation before the experiment. For each QA instance, both models are given the same prompt format, which consists of a system instruction and a user message. The prompt template used for both models in this study is illustrated in Fig 2.

```
System:
You are an educational assistant. Answer the question based solely on the provided
context. Do not use any external knowledge beyond what is given.
User:
Context: [TEXTBOOK PASSAGE]
Question: [QUESTION]
```

Fig 2. Prompt template used for context-grounded question generation

The generation parameters for both models are also kept consistent throughout the experiment, as shown in Table 2. A zero-shot prompting approach is chosen to measure each model's ability to produce contextually relevant answers without prior task examples. The temperature parameter is set to 0.3 to encourage focused and consistent responses while still allowing slight natural variation in language. The maximum token length is limited to 512 to keep the responses concise and suitable for secondary school educational use. In addition, the top-p value is set to 0.95 to apply nucleus sampling, which helps maintain a balanced and controlled range of vocabulary choices during text generation. Overall, these parameter settings are selected to balance response consistency and output diversity across the dataset.

Table 2. Generation parameter configuration

Parameter	Value
Prompting strategy	Zero-shot
Temperature	0.3
Max Tokens	512
Top-p	0.95

3. Evaluation

The evaluation stage of this study used an LLM-as-a-Judge framework, where DeepSeek acts as an automated evaluator to assess the quality of answers generated by GPT-4o-mini and DeepSeek-v4-flash across the entire dataset. This approach has become increasingly popular in recent NLP research because it provides a scalable and cost-efficient alternative to human evaluation, especially for open-ended text generation tasks. DeepSeek is selected as the judge model for two main reasons. First, the model has strong multilingual understanding capabilities, including support for Bahasa Indonesia, making it suitable for evaluating educational responses in the Indonesian context. Second, DeepSeek offers lower API costs compared to GPT-4, which is important considering the large number of

evaluation requests. Each generated answer is evaluated based on three quality dimensions using a 1-5 Likert scale. The definitions and scoring criteria for these dimensions are presented in Table 3.

Table 3. Aspect of evaluation used in the LLM-as-a-judge framework

Aspect of Evaluation	Definition
Answerability (Maity et al., 2025)	The ability of answer generated remains consistent with the context provided and uses only the available information.
Fluency (Wang et al., 2024)	The grammatical accuracy, readability, and natural flow of the generated text in Bahasa Indonesia, reflecting whether the response is clear, coherent, and easy to understand.
Relevance (Maity et al., 2025)	The ability of the generated answer to provide a direct and appropriate response to the given question.

To ensure consistent and structured evaluations, the judge model received an evaluation prompt containing the textbook passage, the question, and the generated answer for each QA instance. The model is instructed to return its evaluation in a structured JSON format, including an integer score for each dimension and a short explanation for the assigned score. The evaluation temperature was set to 0.3, to improve scoring consistency and reduce random variation between evaluations. If an API request failed or produced an invalid response, the record is assigned a default score of zero and excluded from aggregate calculations to avoid affecting the final reported results.

RESULTS AND DISCUSSION

This chapter presents the results of the comparative evaluation of text generation quality between GPT-4o-mini and DeepSeek-v4-flash using the Informatics QA dataset constructed from Senior High School of Informatics textbooks obtained from SISPBK. The evaluation is conducted automatically using DeepSeek as an LLM-as-a-Judge. Each generated answer is assessed based on three quality aspects such as answerability, fluency, and relevance. All aspects are scored using a 1-5 Likert scale. The results of the experiments are shown in Table 4.

Table 4. Comparative Evaluation Results of GPT-4o-mini and DeepSeek-v4-flash

Aspect of Evaluation	GPT-4o-mini			DeepSeek-v4-flash		
	X	XI	XII	X	XI	XII
Answerability	1.80	2.40	**2.60**	1.40	2.00	**2.20**
Fluency	4.60	**5.00**	**5.00**	4.80	4.00	**5.00**
Relevance	2.40	2.60	**2.80**	2.60	2.80	**3.20**

Table 4 presents the average evaluation scores assigned by DeepSeek as the automated judge for answers generated by GPT-4o-mini and DeepSeek-v4-flash across three quality aspects and three grade levels (X, XI, and XII). Overall, both models showed uneven performance across the evaluated dimensions. Fluency consistently received the highest scores, while answerability and relevance showed greater limitations and indicated more room for improvement.

For the answerability, GPT-4o-mini consistently achieved higher scores than DeepSeek-v4-flash across all grade levels (see Table 4). This trend suggests that textbook materials at higher grade levels may provide more complete and clearly structured contexts, making it easier for the models to generate context-based answers. However, despite this improvement, the overall answerability scores remained relatively low, with all values below 3.00. This indicates that both models still had difficulty generating answers that are fully supported by the provided context alone, which is an important requirement in the closed-book QA setting used in this study.

The fluency aspect produced the highest scores for both models across all grade levels (see Table 4). These results show that both models are highly capable of generating grammatically correct and

natural-sounding text in Bahasa Indonesia. This is in line with previous multilingual LLM studies, which suggest that LLM generally perform well in surface-level language generation, even in non-English languages. GPT-4o-mini reached a perfect fluency score of 5.00 for grades XI and XII. In comparison, DeepSeek-v4-flash only achieved a perfect score in grade XII and showed a lower performance at grade XI with a score of 4.00.

For the relevance dimension, both models again showed steady improvement from grade X to grade XII (see Table 4). Overall, DeepSeek-v4-flash consistently outperformed GPT-4o-mini in these aspects across all grade levels, with the largest difference observed at grade XII (see Table 4). This result indicates that although GPT-4o-mini tends to produce more fluent and contextually grounded responses, DeepSeek-v4-flash is somewhat better at aligning its answers with the main topic of the question, even if its responses are less tightly tied to the provided context. Overall, these findings show a clear performance gap across the three evaluation aspects. The consistently high fluency scores stand in strong contrast to the much lower scores for answerability and relevance. This suggests that both models can produce well-written and natural-sounding answers, but they do not always remain strictly grounded in the context provided or fully address the question.

This pattern is especially important in an educational setting, where answerability and relevance are key aspects of answer quality. It highlights the need for better prompt design and possibly further fine-tuning on domain-specific educational materials to improve how well LLM-generated answers stay aligned with curriculum-based QA tasks.

CONCLUSIONS

This study compared the text generation quality of GPT-4o-mini and DeepSeek-v4-flash on context-based QA tasks derived from official Informatics textbooks for Senior High Schools Grade X, XI, and XII sourced from SISPBK. The evaluation is conducted automatically using DeepSeek as an LLM-as-a-Judge, and it assessed three aspects: answerability, fluency, and relevance. The results show that both models performed very well in the fluency dimension, indicating that they can generate grammatically correct and natural-sounding Bahasa Indonesia without task-specific fine-tuning. GPT-4o-mini performed better in answerability, while DeepSeek-v4-flash showed a slight advantage in relevance across all grade levels. However, the relatively low answerability scores for both models consistently below 3.00 suggest that neither model was fully able to meet the requirements of a closed-book, context-limited setting. This highlights a clear gap between general-purpose LLM capabilities and the needs of curriculum-based QA tasks.

In addition, there is a consistent improvement in scores from grade X to grade XII. This suggests that higher-grade textbook content tends to be more structured and provides clearer context, which helps the models generate better answers. Overall, these findings show that while current LLM have strong potential for educational QA applications in Indonesia, their ability to stay grounded in context still needs improvement before they can be reliably used in curriculum-specific settings. Future work should focus on domain-specific fine-tuning using SISPBK materials, improved prompt design, and broader evaluation across different subject areas to enhance the effectiveness of LLM-based systems in Indonesian secondary education.

REFERENCES

Chang, Y., Wang, X. U., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H. A. O., Yi, X., Wang, C., Wang, Y., Ye, W. E. I., Zhang, Y. U. E., Chang, Y. I., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3).

Indarta, Y., Jalinus, N., Samala, A. D., Riyanda, A. R., & Adi, N. H. (2022). Relevansi Kurikulum

Merdeka Belajar dengan Model Pembelajaran Abad 21 dalam Perkembangan Era Society 5 . 0. *EDUKATIF : JURNAL ILMU PENDIDIKAN*, *4*(2), 3011–3024.

Kemendikbudristek. (2024). *Sistem Informasi Perbukuan Indonesia (SIBI)*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. https://buku.kemdikbud.go.id

Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities : Skill tagging of multilingual educational content. *British Journal of Educational Technology*, *December 2023*, 2039–2057. https://doi.org/10.1111/bjet.13465

Maity, S., Deroy, A., & Sarkar, S. (2025). Can large language models meet the challenge of generating school-level questions? *Computers and Education: Artificial Intelligence*, *8*. https://doi.org/10.1016/j.caeai.2025.100370

Sollers, P. (2014). ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew. *Infini*, *127*, 17–21. https://doi.org/10.51644/9781771123907-008

Upadhyay, P., Agarwal, R., Dhiman, S., Sarkar, A., & Chaturvedi, S. (2024). A comprehensive survey on answer generation methods using NLP. *Natural Language Processing Journal*, *8*(June), 100088. https://doi.org/10.1016/j.nlp.2024.100088

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Wang, L., Song, R., Guo, W., & Yang, H. (2024). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2024.2412082

Wentzel, G. (1922). BLEU: a Method for Automatic Evaluation of Machine Translation. *Annalen Der Physik*, *371*(23), 437–461. https://doi.org/10.1002/andp.19223712302

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6). https://doi.org/10.1145/3649506