Engineering and Technology Journal e-ISSN: 2456-3358

Volume 10 Issue 01 January-2025, Page No.- 3559-3565

DOI: 10.47191/etj/v10i01.12, I.F. – 8.227

Enhancing Educational QA Systems: Integrating Knowledge Graphs and Large Language Models for Context-Aware Learning

Ayse Kok Arslan

Researcher, Silcion Valley Chapter-Oxford ALumni

ABSTRACT: This study explores the integration of Knowledge Graphs (KG) and Large Language Models (LLMs) to design a question-answering (QA) system in the field of education. The proposed method involves constructing a KG using LLMs, retrieving contextual prompts from high-quality learning resources, and enhancing these prompts to generate accurate answers to real questions related to major educational concepts.

The technical framework outlined in this paper, along with the analysis of results, contributes to the advancement of LLM applications in educational technology. The findings provide a foundation for developing intelligent, context-aware educational systems that leverage structured knowledge to support learning and enhance educational outcomes.

INTRODUCTION

The emergence of generative large language models (LLMs) has generated significant interest in leveraging their potential to automate interactive question-answering (QA) processes, enabling deeper exploration of conceptual knowledge across diverse subjects.

Initial efforts to integrate LLMs into educational contexts have focused on supporting learner dialogues and delivering feedback through automated text analysis, which facilitates tailored and adaptive learning experiences. These applications span a range of educational tools, including procedural QA systems and individualized learning pathways [9], [10], [13].

Despite these advancements, research on utilizing LLM-based QA systems specifically to enhance learners' conceptual comprehension remains limited. This study addresses this gap by proposing and validating a technical framework designed to develop and implement conceptual QA systems powered by LLMs.

BRIEF OVERVIEW OF LLMS

GenAI technologies such as LLMs (Large Language Models) can analyze the complex patterns and structures of human language and generate human-like text and multimedia content.

LLMs have been described using a wide range of metaphors. While some emphasize its positive potential as a supportive and empowering tool, likening it to a copilot (Risteff, 2023), a sorcerer's apprentice (Liu & Helmer, 2024), a form of co-intelligence (Mollick & Mollick, 2024), or an external brain (Yan et al., 2024), others adopt a more cautious view that acknowledges both its promise and potential risks, describing it as a double-edged sword (Furze, 2024), a kind of magic (Furze, 2024), or a powerful dragon (Bozkurt, 2024a).

On the critical side, some have regarded LLMs as an autotune for knowledge (Cormier, 2023), a colonizing loudspeaker (Gupta et al., 2024), a stochastic parrot (Bender et al., 2021), a dangerous "alien" decision maker (Harari, 2024) or even a weapon of mass destruction (Maas, 2023).

Before delving into their specific use for QA systems, it might be useful to provide an overview of the their general architecture and finetuning mechanism.

The most common ways to finetune language models are instruction finetuning and classification finetuning (Raschka, 2024). Instruction finetuning involves training a language model on a set of tasks using specific instructions to improve its ability to understand and execute tasks described in natural language prompts.

In classification finetuning, the model is trained to recognize a specific set of class labels, such as "spam" and "not spam." Examples of classification tasks extend beyond large language models and email filtering; they include identifying different species of plants from images, categorizing news articles into topics like sports, politics, or technology, and distinguishing between benign and malignant tumors in medical imaging.

The key point is that a classification-finetuned model is restricted to predicting classes it has encountered during its training—for instance, it can determine whether something is "spam" or "not spam", but it can't say anything else about the input text.

Classification finetuning is ideal for projects requiring precise categorization of data into predefined classes, such as sentiment analysis or spam detection (Raschka, 2024).

In contrast to the classification-finetuned model, an instruction-finetuned model typically has the capability to undertake a broader range of tasks. We can view a classification-finetuned model as highly specialized, and generally, it is easier to develop a specialized model than a generalist model that works well across various tasks.

Instruction finetuning improves a model's ability to understand and generate responses based on specific user instructions. Instruction finetuning is best suited for models that need to handle a variety of tasks based on complex user instructions, improving flexibility and interaction quality.

While instruction finetuning is more versatile, it demands larger datasets and greater computational resources to develop models proficient in various tasks. In contrast, classification finetuning requires less data and compute power, but its use is confined to the specific classes on which the model has been trained (Raschka, 2024).

Once a model is initialized with pretrained weights, a modification is required to transform a general pretrained LLM into a specialized LLM for classification tasks. So, it is necessary to modify the pretrained large language model to prepare it for classification finetuning.

It should be taken into account that it's not necessary to finetune all model layers. This is because, in neural network-based language models, the lower layers generally capture basic language structures and semantics that are applicable across a wide range of tasks and datasets. So, finetuning only the last layers (layers near the output), which are more specific to nuanced linguistic patterns and task-specific features, can often be sufficient to adapt the model to new tasks (Raschka, 2024).

A REVIEW OF EXISTING RESEARCH

Recent literature on generative AI (GenAI) presents a dual perspective, mixing enthusiasm with apprehension (Lim et al., 2023; Stracke et al., 2024). This duality arises from GenAI's remarkable ability to process and generate text that rivals human capabilities (Floridi, 2023; Lim et al., 2023; Teubner et al., 2023), prompting some to suggest that it might render the Turing test obsolete.

QA Systems and Learning Applications

A prominent application of Large Language Models (LLMs) is in Question-Answering (QA) systems, which are designed to analyze and interpret natural language questions while retrieving and generating relevant answers. QA systems excel at procedural tasks, offering potential enhancements in learning through guided practice. However, challenges such as hallucination—where LLMs generate plausible but inaccurate responses—remain a significant concern, particularly in educational contexts where accuracy and alignment with curricula are crucial.

Traditionally, QA systems operate by integrating natural language understanding with machine learning to decode user intent, retrieve information from databases, and generate responses through pre-defined templates. Typically, these systems consist of three key modules:

1. **Question Analysis**: Identifying the intent behind the user's query.
2. **Information Retrieval**: Accessing relevant knowledge from databases or knowledge graphs.
3. **Answer Generation**: Producing coherent and contextually appropriate responses, often utilizing LLMs.

QA systems often align with foundational learning theories such as Vygotsky's Zone of Proximal Development (ZPD) and the concept of scaffolding. These frameworks support learners by providing structured assistance, enabling them to tackle problems that lie just beyond their independent capabilities.

Scaffolding, as defined by educational theory, involves temporary support provided by a tutor or system to help learners achieve tasks beyond their current abilities. This concept, closely tied to Vygotsky's ZPD, emphasizes guided learning through four key features:

1. **Shared Understanding**: Aligning goals between the learner and the tutor/system.
2. **Scaffolder Role**: Offering structured guidance by a more knowledgeable agent.
3. **Ongoing Diagnosis**: Continuously assessing learner progress and adapting support accordingly.
4. **Fading**: Gradually reducing assistance as the learner develops competence.

In the context of QA systems, learners engage with AI to address complex problems, advancing their understanding within their ZPD while receiving dynamic, tailored feedback.

Despite their promise, several challenges limit the broader adoption of LLMs in educational QA systems:

1. **Hallucination**: LLMs often produce incorrect yet convincing answers, which can mislead learners and undermine the credibility of their responses (e.g., [12], [14]).
2. **Lack of Domain Knowledge**: While LLMs excel in general contexts, their responses can lack the depth and specificity required for domain-specific teaching and learning ([8], [15]).
3. **Opacity and Explainability**: LLMs function as black-box models, with their internal reasoning and decision-making processes remaining opaque ([16], [17]). This limits the ability to verify or explain their outputs, raising concerns about their reliability.

Moreover, while LLMs provide scaffolding within the ZPD, excessive reliance on them may lead to over-scaffolding, reducing cognitive challenges and hindering the learner's autonomy and intellectual growth.

Retrieval-Augmented Generation (RAG): A Promising Solution

To address these limitations, researchers have explored retrieval-augmented generation (RAG) methods. RAG integrates external knowledge sources—such as syllabi, workbooks, or knowledge graphs (KGs)—into LLM workflows to enhance accuracy and contextual relevance. During inference, RAG incorporates retrieved knowledge into prompts, improving the quality and groundedness of the generated content.

A potential enhancement to traditional RAG methods is the use of knowledge graphs to create semantic networks of interrelated concepts. By mapping relationships between knowledge entities, KGs provide clear, interconnected views of learning content, offering both depth and precision in addressing learner queries.

The effectiveness of RAG in educational contexts can be evaluated through two primary perspectives:

1. **Semantic Similarity**: Measuring the alignment between the retrieved knowledge and the context of the generated response.
2. **Contextual Groundedness**: Assessing the relevance and coherence of the generated content within the learning domain.

By leveraging these frameworks, researchers aim to optimize QA systems, ensuring that LLMs serve as reliable, effective tools for enhancing learning experiences while mitigating their inherent limitations.

QA System Based on RAG and LLM

The integration of Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) in Question-Answering (QA) systems represents a transformative approach to embedding disciplinary knowledge into intelligent question-response mechanisms. This integration significantly enhances learners' efficiency in self-directed study by optimizing the limited availability of their time and cognitive resources. Notably, LLMs have been extensively employed in program tutoring and problem-solving applications, including automated question-answering systems and knowledge recommendation platforms [9], [10].

State-of-the-art natural language processing (NLP) models, such as BERT, Llama, and GPT, are foundational in the development of open-domain chatbots. These models, initially trained on expansive general-domain datasets, can be fine-tuned to specific tasks, thereby aligning more closely with the nuanced requirements of their intended applications [8], [3]. Among these, advanced LLMs like GPT-4 and Llama-2 and -3 have garnered widespread recognition for their ability to generate coherent, contextually appropriate text, underlining their profound impact on NLP advancements.

From a theoretical perspective, RAG enhances the reliability of text generated by LLMs by retrieving information from external knowledge sources. This reliability—often referred to as "faithfulness"—is assessed through the semantic correlation between the LLM-generated output and accurate reference texts, with comparisons frequently drawn against responses provided by human tutors. Empirical evidence suggests that RAG can significantly improve the performance of LLM-based QA systems in educational contexts.

For instance, a QA system utilizing the Llama-3 model, which applies cosine similarity within the RAG framework for text vector retrieval, demonstrated a 9.84% increase in accuracy on a test set comprising non-graphical multiple-choice questions across various STEM disciplines [8].

The RAG framework is typically implemented using one of three primary methodologies:

1. **Template-Based Retrieval**: This approach employs keyword-matching algorithms to identify and retrieve relevant textual materials [2].
2. **Semantic Vector Retrieval**: By extracting high-dimensional matrices (semantic vectors) through NLP models, this widely adopted method leverages its simplicity and effectiveness for similarity-based retrieval [8], [3].
3. **Knowledge Graph-Based Retrieval**: Distinct from text-based retrieval, this method organizes data in a graph structure, forming a knowledge base represented as interconnected nodes (entities) and edges (relations). Subgraphs are retrieved using semantic vector-based methods [2].

Knowledge graphs (KGs) are constructed through the assembly of "triples," in which a head entity is linked to a tail entity via a predicate that defines their relationship [12]. These triples coalesce into a multi-graph framework, with nodes representing entities and edges corresponding to relationships. NLP techniques facilitate the extraction of knowledge entities and their interrelations from extensive educational resources, enabling the creation of knowledge graphs as external knowledge repositories.

The construction of KGs in the educational domain has evolved from early rule-based and lexical methods to modern machine learning and deep learning-based approaches. While lexical and rule-based techniques rely on manual rule formulation by domain experts—limiting scalability—contemporary methods leverage statistical models and neural architectures to manage vast datasets efficiently. For example, convolutional neural networks (CNNs) have been used to extract KGs from MOOC resources, including curricula and textbooks, to support instructional design and provide tailored course recommendations [5].

Knowledge graphs further enable the elucidation of complex concepts by linking foundational ideas to advanced topics through structured entity relationships. This interconnected representation facilitates a progressive learning experience, allowing students to comprehend sophisticated subjects with greater ease [4].

Once a knowledge graph is integrated into a QA system, the system can accurately respond to user queries by analyzing the questions and retrieving relevant entities and their interconnections from the knowledge base. The process involves extracting keywords, entities, and relationships from textual documents, often guided by LLM-generated prompts. These entities and relationships are converted into text embeddings, which are then used to search for related entities. If the similarity threshold (e.g., a cosine similarity of 0.8) is unmet or the entity appears for the first time, separate subgraphs are generated to maintain coherence and integrity within the graph.

This workflow, illustrated in Figure 1, underscores the pivotal role of KG construction and retrieval in enhancing QA system capabilities, thereby advancing the intersection of NLP, knowledge representation, and educational technology.
4o

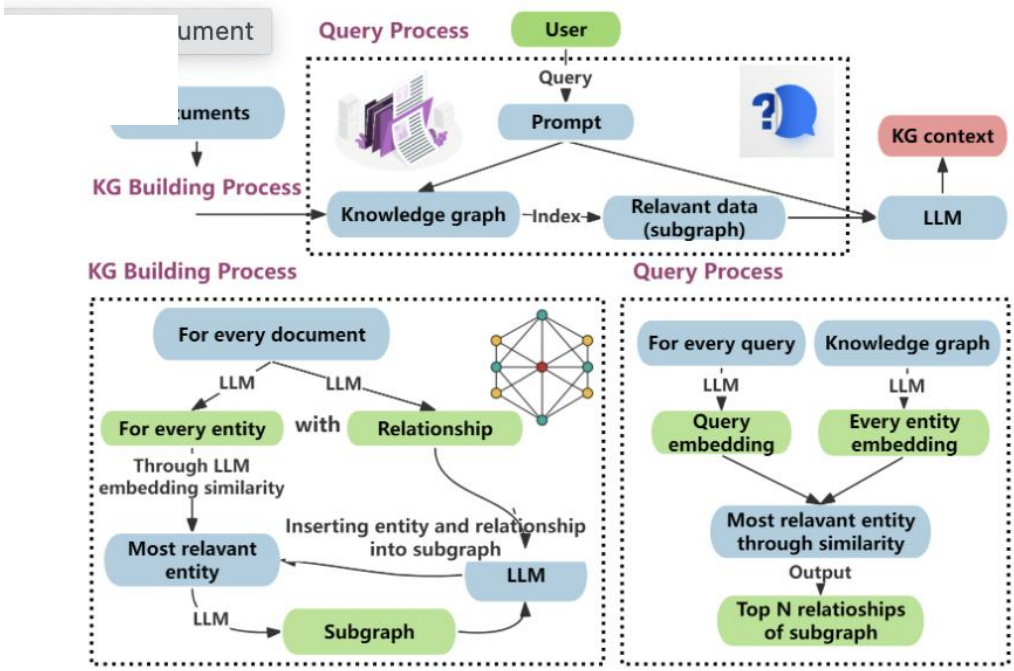

Figure 1. KG Workflow

Relevant relationships within the knowledge graph (KG) are incorporated into the subgraph of the target entity based on the required relational context as determined by the Large Language Model (LLM). At this stage, the LLM primarily oversees the merging of entities and relationships. Each subgraph comprises multiple foundational relationships, with each relationship defined as a pair of entities and their associated connection. The process of merging entities and relationships relies on well-established NLP disambiguation and extraction algorithms, which are fundamental tasks in the domain of computer science.

In the retrieval query workflow, illustrated in the bottom-right corner of the diagram, the learner's query is processed to compute cosine similarity between its semantic vector and the semantic vectors of each entity in the KG. This step identifies the entity most closely aligned with the query. Subsequently, the system retrieves the most relevant relationships within the subgraph of the identified entity. This is achieved by evaluating the similarity between the semantic vectors of entity pairs and relationships and the query vector, using a predefined similarity threshold. The extracted relationships are then concatenated to form a coherent context, which is input into the LLM to generate the KG-based response.

Finally, the learner's query and the contextual output from the KG (i.e., the retrieved information) are jointly fed into the LLM. The LLM synthesizes this information to produce a comprehensive and contextually accurate response to the learner's question, ensuring that the answer is both relevant and grounded in the structured knowledge within the graph.

RECOMMENDATIONS

The integration of Generative AI (GenAI) into the data-to-wisdom continuum marks a transformative shift in the dynamics of human wisdom, with synthetic information emerging as a significant counterpart to human-generated organic information. While GenAI holds the potential to expand cognitive horizons and facilitate complex tasks, its role must be carefully delineated to avoid fostering over-dependence. The balance between support and over-reliance is particularly critical in educational contexts, where the objective is to cultivate independent intellectual growth alongside technological augmentation.

GenAI's capacity to revolutionize interdisciplinary learning lies in its ability to synthesize and integrate knowledge across domains. However, such potential can only be realized if learners are guided toward critical synthesis, rather than

passive consumption of AI-generated outputs. Critical reflection is indispensable; without it, learners risk accepting AI-derived connections at face value, thereby undermining the deeper understanding that interdisciplinary learning demands

CONCLUSION

This study investigates the application of a Knowledge Graph (KG)-based Retrieval-Augmented Generation (RAG) question-answering (QA) system, which incorporates retrieval enhancement to produce high-quality responses tailored for training in conceptual learning contexts.

The technical framework and the analytical findings presented herein offer valuable contributions to the fields of conversational AI, intelligent tutoring systems, and Large Language Model (LLM)-related research.

Looking ahead, future research could explore the development of a hybrid model that effectively merges the natural language understanding capabilities of logical reasoning models with the structured knowledge representation inherent in knowledge bases. Such a model would empower educational systems to not only generate human-like text but also retrieve and reason with structured knowledge, thereby advancing the development of more intelligent, context-aware educational assistants.

REFERENCES

1. Ansari, A. N., Ahmad, S., & Bhutta, S. M. (2024). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. Education and Information Technologies, 29(9),1128111321. https://doi.org/10.1007/s10639-023-12223-4
2. Bayne, S., Evans, P., Ewins, R., Knox, J., Lamb, J., Macleod, H., O'Shea, C., Ross, J., Sheail, P., Sinclair, C., & Johnston, K. (2020). The manifesto for teaching online. MIT Press. https://doi.org/10.7551/mitpress/11840.001.0001
3. Bayne, S., & Ross, J. (2016). Manifesto redux: Making a teaching philosophy from networked learning research. In S. Cranmer, N. B. Dohn, M. de Laat, T. Ryberg, & J. A. Sime (Eds.), Proceedings of the 10th International Conference on Networked Learning 2016 (pp. 210–128). Lancaster: University of Lancaster. https://doi.org/10.54337/nlc.v10.8821
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922
5. Birt, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. M. (2016). Member checking: A tool to enhance trustworthiness or merely a nod to validation? Qualitative Health Research, 26(13), 1802-1811. https://doi.org/10.1177/1049732316654870
6. Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., Farrow, R., Bond, M., Nerantzi, C., Honeychurch, S., Bali, M., Dron, J., Mir, K., Stewart, B., Costello, E., Mason, J., Stracke, C. M., Romero-Hall, E., Koutropoulos, A., Toquero, C. M., Singh, L., Tlili, A., Lee, K., Nichols, M., Ossiannilsson, E., Brown, M., Irvine, V., Raffaghelli, J. E., Santos-Hermosa, G. Farrell, O., Adam, T., Thong, Y. L., Sani-Bozkurt, S., Sharma, R. C., Hrastinski, S., & Jandrić, P. (2023a). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. Asian Journal of Distance Education, 18(1), 53–130 https://doi.org/10.5281/zenodo.7636568
7. Burns, T., Sinfield, S., & Abegglen, S. (2023). Postdigital academic writing. In Encyclopedia of Postdigital Science and Education (pp. 1–7). Springer Nature Switzerland https://doi.org/10.1007/978-3-031-35469-4_27-1
8. Clayton, M. J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. Educational Psychology, 17(4), 373–386 https://doi.org/10.1080/0144341970170401
9. Cormier, D. (2023, Jan 20). ChatGPT search – Autotune for knowledge. Dave's Educational Blog. https://davecormier.com/edblog/2023/01/20/chatgpt-search-autotune-for-knowledge/comment-page-1/
10. Costello, E. (2024). ChatGPT and the educational AI chatter: Full of bullshit or trying to tell us something? Postdigital Science and Education, 6(2), 425–430. https://doi.org/10.1007/s42438-023-00398-5
11. Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. Philosophy & Technology, 36(1),1–7. https://doi.org/10.1007/s13347-023-00621-y
12. Furze, L. (2024, July 19). AI metaphors we live by: The language of artificial intelligence. https://leonfurze.com/2024/07/19/ai-metaphors-we-live-by-the-language-of-artificial-intelligence/
13. Gale, K., & Bowstead, H. (2013). Deleuze and collaborative writing as method of inquiry. Journal of Learning Development in Higher Education, 6. https://doi.org/10.47408/jldhe.v0i6.222
14. Gupta, A., Atef, Y., Mills, A., & Bali, M. (2024). Assistant, parrot, or colonizing loudspeaker? ChatGPT metaphors for developing critical AI literacies. Open Praxis, 16(1), 37–53. https://doi.org/10.55982/openpraxis.16.1.631

15. Harari, Y. N. (2024). Nexus: A brief history of information networks from the stone age to AI. Random House Publishing Group. https://www.ynharari.com/book/nexus/
16. Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. Ethics and Information Technology, 26(2), 38. https://doi.org/10.1007/s10676-024-09775-5
17. Jandrić, P., Hayes, D., Truelove, I., Levinson, P., Mayo, P., Ryberg, T., Monzó, L. D., Allen, Q., Stewart, P. A., Carr, P. R., Jackson, L., Bridges, S., Escaño, C., Grauslund, D., Mañero, J., Lukoko, H. O., Bryant, P., Fuentes-Martinez, A., Gibbons, A., … Hayes, S. (2020). Teaching in the Age of Covid-19. Postdigital Science and Education, 2(3), 1069–1230. https://doi.org/10.1007/s42438-020-00169-6
18. Jandrić, P., Luke, T. W., Sturm, S., McLaren, P., Jackson, L., MacKenzie, A., Tesar, M., Stewart, G. T., Roberts, P., Abegglen, S., Burns, T., Sinfield, S., Hayes, S., Jaldemark, J., Peters, M. A., Sinclair, C., & Gibbons, A. (2023). Collective writing: The continuous struggle for meaning-making. Postdigital Science and Education, 5(3), 851–893. https://doi.org/10.1007/s42438-022-00320-5
19. Koutropoulos, A., Stewart, B., Singh, L., Sinfield, S., Burns, T., Abegglen, S., Hamon, K., Honeychurch, S., & Bozkurt, A. (2024). Lines of flight: The digital fragmenting of educational networks. Journal of Interactive Media in Education, 2024(1), 1–13 https://doi.org/10.5334/jime.850
20. Latour, B. (2010). An attempt at a "compositionist manifesto". New Literary History, 41(3), 471–490. https://doi.org/10.1353/nlh.2010.a408295
21. Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The International Journal of Management Education, 21(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790
22. Liu, D., & Helmer, E. (2024, Feb 9). The sorcerer's apprentice: Applied AI for data. Perspectives, Substack. https://debliu.substack.com/p/the-sorcerers-apprentice-applied
23. Maas, M. M. (2023). AI is like… A literature review of AI metaphors and why they matter for policy. SSRN Electronic Journal https://doi.org/10.2139/ssrn.4612468
24. Mack, E. (2014, Oct 26). Elon Musk: 'We are summoning the demon' with artificial intelligence. CNET. https://www.cnet.com/science/elon-musk-we-are-summoning-the-demon-with-artificial-intelligence/
25. Merriam, S. B. (2001). Qualitative research and case study applications in education. Jossey-Bass.
26. Mollick, E., & Mollick, E. (2024). Co-Intelligence. Random House UK.
27. Pelletier, K., Brown, M., Brooks, D. C., McCormack, M., Reeves, J., Arbino, N., Bozkurt, A., Crawford, S., Czerniewicz, L., Gibson, R., Linder, K., Mason, J., & Mondelli, V. (2021). 2021 EDUCAUSE horizon report teaching and learning edition.EDUCAUSE. https://www.learntechlib.org/p/219489/
28. Peters, M. A., Besley, T., Tesar, M., Jackson, L., Jandrić, P., Arndt, S., & Sturm, S. (2021). The methodology and philosophy of collective writing: An educational philosophy and theory reader volume X. Routledge. https://doi.org/10.4324/9781003171959
29. Risteff, M. (2023, Apr 24). AI is an extraordinary copilot when used responsibly. K-12 Dive. https://www.k12dive.com/spons/ai-is-an-extraordinary-copilot-when-used-responsibly/647644/
30. Saban, A., Kocbeker, B. N., & Saban, A. (2007). Prospective teachers' conceptions of teaching and learning revealed through metaphor analysis. Learning and Instruction, 17(2), 123–139. https://doi.org/10.1016/j.learninstruc.2007.01.003
31. Sharples, M., & Pérez y Pérez, R. (2022). Story machines: How computers have become creative writers. Routledge. https://doi.org/10.4324/9781003161431 Slagter van Tryon, P. J., & Bishop, M. J. (2006). Identifying "e-mmediacy strategies for web-based instruction: A Delphi study." The Quarterly Review of Distance Education, 7(1) 49–62.
32. Stracke, C. M., Burgos, D., Santos-Hermosa, G., Bozkurt, A., Sharma, R. C., Swiatek Cassafieres, C., dos Santos, A. I., Mason, J., Ossiannilsson, E., Shon, J. G., Wan, M., Obiageli Agbu, J.-F., Farrow, R., Karakaya, Ö., Nerantzi, C., Ramírez-Montoya, M. S., Conole, G., Cox, G., & Truong, V. (2022a). Responding to the initial challenge of the COVID-19 pandemic: Analysis of international responses and impact in school and higher education. Sustainability, 14(3), 1876. https://doi.org/10.3390/su14031876
33. Stracke, C. M., Chounta, I.-A., & Holmes, W. (2024). Global trends in scientific debates on trustworthy and ethical artificial intelligence and education. Artificial Intelligence in Education. Communications in Computer and Information Science, 2150, 254–262 https://doi.org/10.1007/978-3-031-64315-6_21

34. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10(1), 1–24 https://doi.org/10.1186/s40561-023-00237-x
35. Yan, Y., Sun, W., & Zhao, X. (2024). Metaphorical conceptualizations of generative artificial intelligence use by Chinese university EFL learners. Frontiers in Education, 9 https://doi.org/10.3389/feduc.2024.1430494
36. Ziglio, E. (1996). The Delphi method and its contribution to decision-making. In M. Adler & E. Ziglio (Eds.), Gazing into the oracle: The Delphi method and its application to social policy and public health (pp. 3–33). Jessica Kingsley Publishers