Lessons from a Multimodal and Trustworthy AI System for Intelligent Textbooks

Karan Taneja[1], Anjali Singh[2] and Ashok K. Goel[1]

[1]Georgia Institute of Technology, Atlanta GA 30332, USA
[2]University of Texas at Austin, Austin TX 78712, USA

Abstract
We present an experimental study with MuDoC, a Multimodal Document-grounded Conversational AI system built to improve learner experience with digital textbooks. Going beyond text-based systems, MuDoC leverages grounded visuals in its interleaved text-and-image responses, and allows seamless navigation within the textbook to examine the source content. We conducted a user study where learners solved analytical problems using MuDoC and a baseline text-only system. Through surveys and post-study interviews, we examined user experience, aiming to measure the impact of multimodality and content verifiability on learner engagement and trust. The results show the potential of MuDoC as an educational tool in promoting deeper engagement with textbooks through multimedia learning and by lowering the barrier to verifying AI-generated information.

Keywords
Multimodal AI, Source Attribution, Trustworthy AI, Learner Engagement

1. Introduction

Conversational AI has the potential to act as a powerful medium for interacting with digital textbooks by grounding its responses in relevant content retrieved from the textbook. The process of interacting with textbooks in this manner enables opportunities for personalized tutoring through structured and reliable sources of information, unlike AI responses that are generated using general web-based information. Previous work [1, 2] has explored the use of text-based systems that answer queries based on educational content to support students in course-specific information seeking. Building on the success of these systems, we recently proposed MuDoC [3, 4], a Multimodal Document-grounded Conversational AI system, that further leverages visuals grounded in textbooks to generate multimodal responses with interleaved text and images. Further, MuDoC allows seamless navigation across the AI-generated responses and the textbook to examine the relevant textbook content by clicking on paragraphs and images in the AI-generated responses.

In this paper, we present an experimental study where we invited n=30 graduate students to solve problems, based on a textbook, using MuDoC versus a baseline system which is its simpler text-only counterpart. Through surveys and post-study interviews, we examine the differences in user experience along with pros and cons of different features in supporting learning. The goal of the study was to measure the impact of MuDoC on engagement and trustworthiness owing to the multimodality and verifiability of its responses.

The qualitative feedback reveals the potential of MuDoC as an educational tool that promotes deeper engagement with textbooks. With higher interactivity and trust, MuDoC encouraged learners to read the textbook to better understand the context of text and images in AI responses, build coherence between the two modalities, and identify limitations of AI responses. We conclude with insights for improving multimodal generative AI systems for interacting with textbooks.

iTextbooks'25: Sixth Workshop on Intelligent Textbooks, July 26, 2025, Palermo, Italy
karan.taneja@cc.gatech.edu (K. Taneja); anjali.singh@ischool.utexas.edu (A. Singh); ashok.goel@cc.gatech.edu (A. K. Goel)
https://krntneja.github.io/ (K. Taneja); https://sites.google.com/view/singhanj13 (A. Singh); https://dilab.gatech.edu/ashok-k-goel/ (A. K. Goel)
0000-0002-7921-8795 (K. Taneja); 0009-0006-4092-4779 (A. Singh); 0000-0003-4043-0614 (A. K. Goel)

2. Related Work

Jill Watson [1] is a document-grounded conversational AI agent based on GPT-3.5 and GPT-4. It leverages retrieval models and LLMs for retrieval-augmented generation (RAG) to answer queries related to the document provided by course instructors. Pedagogical Tutor (PET) [2] is a dialog system that relies on template-based slides for document structure and answers student questions using GPT-4. Further, its chat window within the slide viewer allows students to search for relevant slides. Intelligent textbooks based on Jill Watson, proposed by Olson et al. [5], similarly display a chat window next to the textbook but does not leverage visuals. Text-based document-grounded dialog [6], visually-grounded chat [7], and multimodal document understanding [8, 9] have also been explored in other works.

Curio [10] processes and generates explanations for diverse content types within textbooks, and integrates them with a video player to deliver these explanations and recommendations. Phygital Textbook [11] integrates a physical textbook with a supplementary digital interactive interface. The digital layer incorporates quizzes, AR/3D models and animations, audio narration, and external links to support learners engaging in multiple modes of interaction. Our system, MuDoC [3], uses a pipeline similar to Jill Watson, but additionally extracts figures from documents during the preprocessing step to utilize them for generating interleaved text and image responses. Further, its interface can be used to jump to the source in the textbook to verify AI-generated information.

Cognitive Theory of Multimedia Learning [12] suggests that text and visuals together lead to better learning gains compared to text-only content, which motivates the development of MuDoC as an AI-powered interface for intelligent textbooks. MuDoC also provides source attribution and quick navigation, which can improve trust in AI systems [13]. From the perspective of the iTextbooks Pedagogical Framework [14], MuDoC utilizes the strategies of multimedia learning, adaptive learning, and personalized learning to create an engaging and effective learning environment.

3. MuDoC and Study Design

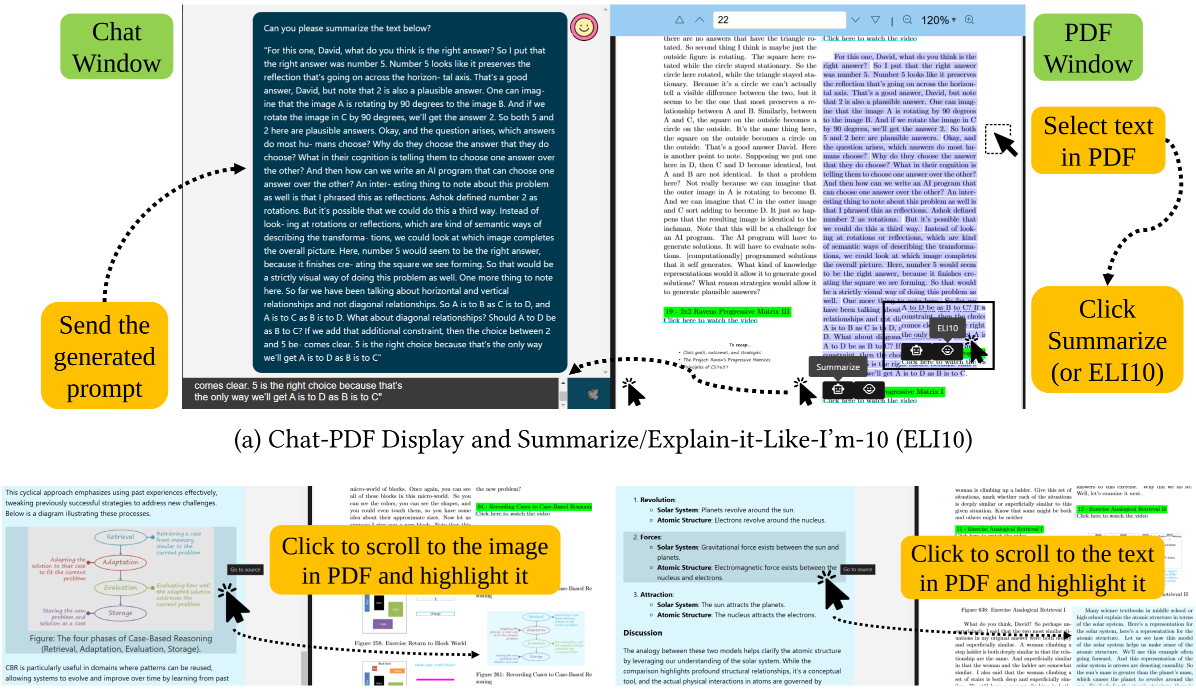
MuDoC [3] provides an intelligent textbook experience through its user interface (UI) designed to foster improved engagement with textbook content and enhance the verifiability of information for higher trust. It is primarily targeted for higher education, aiming to be integrated into classrooms as a tool within Learning Management Systems (LMS). A video demonstration of MuDoC available here[1].

3.1. User Interface Features

Summarize and Explain-it-Like-I'm-10 (ELI10): As illustrated in Figure 1a, the user interface presents a simple layout: a chat window occupies the left side of the display, featuring a text-box at the bottom for users to input queries, while a PDF textbook is prominently displayed on the right. MuDoC supports reading difficult texts through its "Summarize" and "ELI10" (Explain-it-Like-I'm-10) features. To utilize these, a user can simply drag their cursor to select a paragraph within the PDF. Upon selection, 'Summarize' and 'ELI10' options become visible. Selecting an option generates a corresponding prompt, which is then editable in the chat text-box. This allows the user to edit and refine the prompt, such as by modifying the prompt to specify desired length, additional context, or even asking for related images. The "Summarize" feature yields a concise version of the selected text, facilitating quick reading and comprehension, while the "ELI10" feature provides an explanation in simpler terms, significantly improving the readability of convoluted or unfamiliar content. The goal of these is to make the process of engaging with long and intricate textbooks more dynamic and appealing for learners.

Navigation Using Images and Texts: MuDoC enhances content verifiability and deeper learning by enabling seamless navigation between AI-generated responses and the textbook. When a user hovers their cursor over an image or a paragraph presented in a multimodal AI response, the element

[1] https://www.youtube.com/watch?v=yCx_2PXXhO8



(a) Chat-PDF Display and Summarize/Explain-it-Like-I'm-10 (ELI10)

(b) Clickable Images for Navigation

(c) Clickable Text for Navigation

Figure 1: User Interface Design and Features: (a) Users can select the text in PDF document and select either 'Summarize' or 'ELI10' to create a prompt that is sent to the AI system. (b) and (c) Figures and paragraphs in AI responses are clickable and allow users to seamlessly navigate to the source in the document which is highlighted for a few seconds.

visually indicates its clickability through a subtle brightness change and the appearance of tooltip text, as depicted in Figure 1b (for images) and Figure 1c (for text). When an image or a paragraph is clicked, the PDF view automatically scrolls to the correct page in the textbook where the corresponding figure appears or a similar text snippet is found. For visual clarity and immediate context, the figure or the matched text is then highlighted for three seconds before fading away. For images, the system leverages pre-processed bounding box information, which includes the image's page number, page size, location, and dimensions. This metadata, derived from document layout analysis during preprocessing, allows for precise navigation. For paragraphs, since LLM-generated text is typically rephrased and not verbatim from the source, a post-processing step is performed after response generation. In this step, paragraphs in the AI response are mapped to raw text snippets in the textbook based on the cosine similarity between their DPR context embeddings, which can sometimes lead to imperfect mappings. This navigation capability using images and paragraphs allows users to examine the original context in which a figure or text appears within the textbook. It encourages them to delve deeper by reading from the textbook, which serves as a more complete source of knowledge in contrast to AI-generated text that, while helpful, can sometimes suffer from hallucination or a lack of granular context.

3.2. Study Design

We recruited n=30 graduate students from Georgia Institute of Technology (Atlanta, GA, USA) with ML/AI backgrounds, but new to the course on Knowledge-based AI upon which the textbook [15] is based. We compared MuDoC to a baseline system called TexDoC, a simpler version of MuDoC that provides text-only responses without source attribution. Each participant solved problems related to Analogical Reasoning (ANA) and Incremental Concept Learning (INC) in a random order using the two systems, each within a 20-minute time limit. They were instructed to learn the concepts relevant to the problem by posing questions to the AI system, and never to directly ask the AI to solve the problem. Surveys and post-study interviews were used to gather quantitative and qualitative feedback.

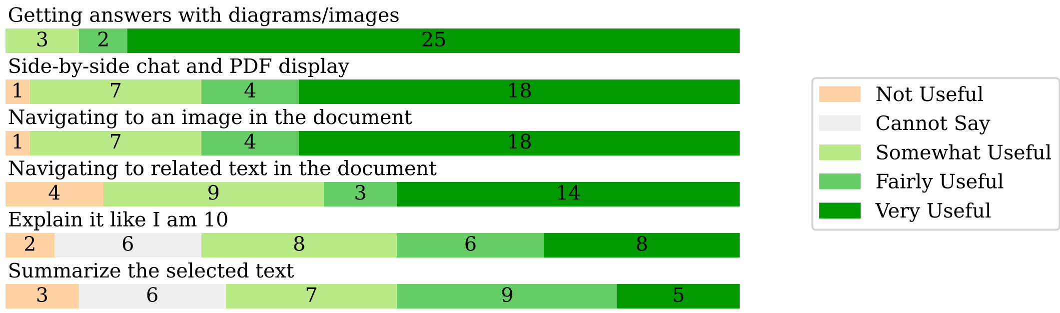


Figure 2: Perceived 'usefulness' of different MuDoC features sorted by aggregated preferences. Numbers indicate participant count. Users could select 'Cannot Say' when they did not employ a feature.

4. Results

The participants in the study asked M=5.17, SD=2.40 questions to MuDoC and M=5.86, SD=2.68 questions to TexDoC, but a paired t-test revealed that this difference was not significant (t=1.481, p=0.149). They spent an average of 19 minutes with both systems, close to the 20 minutes limit. An average response had 290 words (SD=10) and MuDoC responses additionally contained 1-5 images from the textbook (M=2.02, SD=1.18). The mean time-to-first-token for TexDoC and MuDoC was 4.1s and 6.9s respectively.

4.1. UI Features and Feedback

We asked participants to rate the perceived usefulness of six UI features of MuDoC on a 5-point Likert scale. The results are presented in Fig. 2. We also asked participants for qualitative feedback through questions about their experience with both the systems. We now report these results and discuss the strengths and limitations highlighted by the participants.

Multimodality: All 30 participants found MuDoC *responses with images* to be useful. Most believed that visuals were more effective in explaining concepts than text-only responses were. One participant elaborated, *"Visual examples help a lot. For text, you have to mentally map it or write it while reading."*. Another participant said, *"Looking at diagrams helped create the diagrams,"* because some sub-tasks required creating diagrams. One participant pointed out correctly that images could not be hallucinated, as they were clipped from the textbook, highlighting the increase in trust owed to grounded images. Many believed that images made responses more engaging, interesting and memorable. Many students went as far as to identify themselves as visual learners, unaware of the overwhelming evidence against the myth of learning styles [16]. Overall, these findings highlight the need for multimodality in conversational AI for intelligent textbooks. Besides these strengths, a few limitations were also identified. Some participants felt overwhelmed by multiple images and found the responses with images to be verbose. One participant said, *"I felt as though [MuDoC] provided more images than were necessary"*. A few participants also expressed concerns about imperfections in images as some of them were clipped short/large and consecutive figures were occasionally merged into one large figure.

Side-by-side Chat and PDF: 29 participants preferred the convenience of *side-by-side chat and PDF display*, mainly because it enabled verification and increased trust. Many explained that it helped them trust the AI response when they could confirm it by reading from the book. One participant said, *"I would prefer using [MuDoC] because, with the book, I feel the knowledge is more credible."* However, on the flip side, one participant commented *"I like the images but I don't care about the textbook. Constantly having the book was not as helpful."*

Navigation using Images and Text: 29 and 26 participants identified *navigation using images and text* respectively as useful due to the convenience in referring to the textbook, finding relevant content, and providing credibility to responses. One participant stated, *"With [MuDoC], making an effort to read the textbook more often was easy."* Another said, *"I'm a fan of reading books and the option to navigate in the book is just better, simply believing [AI responses] or manually finding [similar content] is harder"*. Many participants also regarded navigation as useful because it made the AI responses more credible. One participant explained, *"It showed where [the content] was in the textbook, so I knew [MuDoC] was not hallucinating"*. However, some expressed their frustration with navigating using text, as a few paragraphs in the AI responses were mapped to headings instead of text or to snippets that were not related. One participant elaborated, *"When clicking on the text to go to the source, it would instead send me to a non-relevant location in the text and I would have to manually scroll to the relevant sections."* Two participants also expressed concern about low contrast of the highlighted box, especially when the mapped text had a non-white background color.

ELI10 and Summarize: 22 and 21 participants found ELI10 and Summarize features to be useful, respectively. Many found them to be engaging and assistive as they could shorten or simplify the text from textbook, while others could not spare time to utilize them because of the time constraints.

4.2. Other Observations

Learner Engagement: The participants found MuDoC to be very engaging as it supported *interactive exploration* of the textbook. One participant stated, *"[MuDoC elements] are interactive—you're not just staring at content; you're navigating through it. It makes me want to spend more time with [MuDoC]."* Some participants found the markdown-rich format of responses to be *instructive and interesting*. One participant explained, *"I liked the format of the answers, with the definitions and diagrams and bolded text—this is also how I take notes!"* Some also indicated that MuDoC led to *deeper engagement* compared to TexDoC. One participant said, *"[MuDoC] made me think more about the [AI] answer and check if it was true. If there was a text or diagram that wasn't a part of the answer, it made me wonder why it wasn't included. And when hallucinations are there, it lowers the barrier to verifying the text and being more cautious."* Participants also placed *higher trust* in MuDoC because of the verifiability of the content based on the textbook. However, one participant referring to the quality of retrieved images said, *"Trustworthiness comes from concise and direct answers, [MuDoC] seemed to be covering some bases and just gave too much information."* This suggests that we need to improve the visual retrieval pipeline and ensure that images are consistently relevant to the user query and their learning needs.

Impact of Problem Difficulty: Participants rated the difficulty of ANA and INC problems on a scale of 1 to 10. ANA (M=6.67, SD=1.19) was considered significantly more difficult (t=4.668, p<0.001) than INC (M=4.83, SD=1.53) which explains a higher number of questions asked (t=2.469, p=0.019) for solving ANA problem (M=6.07, SD=2.48) compared to INC (M=4.97, SD=2.55) and more time spent (t=3.737, p<0.001) on solving ANA problem (M=20.47, SD=2.7) compared to INC (M=18.50, SD=2.7). To understand the impact of problem difficulty on user experience, we compared the response distribution using Mann-Whitney U test. When participants solved the easier problem (INC) using MuDoC, it led to favorable ratings for MuDoC (U=66.0, p=0.047). A few participants also acknowledged that difficulty levels of the two problems may have impacted their opinion of the assigned system.

4.3. Discussion and Future Work

The multimodal features of MuDoC led to increased engagement among learners, but the quality of its responses can be further improved by improving the retrieval pipeline, especially for visual information. The current implementation uses a single-step embedding-based retrieval process. Some recent work suggests adding a re-ranking step after retrieval can improve the overall retrieval quality [17, 18]. In the development of MuDoC 2.0, we plan to use an additional re-ranking step for both text and images to improve the relevance of the content used for response generation.

Beyond improving relevance of retrieved content, the response generation step can be improved by relying on principles of Cognitive Theory of Multimedia Learning (CTML) [12]. CTML suggests that multimedia content should contain complementary visual and verbal information that enhances the other modality with minimal redundancy. However, generating such content requires deeper understanding and reflection on coherence between the two modalities. Chain-of-thought reasoning and 'thinking' models can be used to perform complex reasoning tasks [19] to reflect on the retrieved content before generating a response that succinctly answers the query with greater coherence and lower redundancy between the two modalities. Further, few-shot prompting or in-context learning can improve performance over zero-shot prompting [20] by contextualizing the query with examples of desired characteristics in the responses. While these techniques can increase latency or aggravate prompt length constraints, it may prove to be a worthwhile trade-off for the improvement in performance.

MuDoC's main objective is to provide the most relevant multimodal information based on a textbook, but it further allows verification of content through accessible source attribution, which supports increased trust in the system. Reliability i.e. consistent high quality, and credibility i.e. believability of the source are among the most important predictors of trust in information seeking applications [13]. Based on the feedback from participants, MuDoC's reliability can be further increased by improving source attribution quality. We plan to explore hybrid similarity methods [21] for source attribution in the future.

Finally, based on the feedback on MuDoC features, we found that it will be useful to extend *Summarize* and *ELI10* features to implement a drag-to-select functionality for image crops from the textbook, allowing users to include the selected image as part of the prompts. We also observed that multiple sources of information in MuDoC can overwhelm users through information overload. According to cognitive load theory [22], this occurs when the information processed by users exceeds their working memory capacity. To address this issue, MuDoC 2.0 will allow additional controls for users to optionally hide the document, and minimize text and images in the responses. Additionally, incorporating features that allow users to delve deeper into specific parts of the AI response—by elaborating or simplifying the text using LLMs—can help manage cognitive load and improve user experience.

5. Conclusion

This paper presented a study examining the impact of different features of MuDoC designed to enhance intelligent textbook interaction by improving engagement and trust through multimodality and content verifiability. The study compared MuDoC, a Multimodal Document-grounded Conversational AI system, with a text-only baseline and demonstrated the value of integrating multimodal elements and navigation features in AI-powered educational tools. User feedback highlighted the effectiveness of multimodal responses, particularly those with visuals, in explaining concepts, boosting engagement, and building trust due to their grounded nature. The side-by-side chat and textbook display, along with the ability to navigate directly to the source content, significantly enhanced credibility and convenience for users. While features like "Summarize" and "Explain-it-Like-I'm-10" also proved beneficial for text simplification, the study identified limitations such as verbose responses, image clipping imperfections, navigation inaccuracies, and potential familiarity bias with text-only interfaces. Future work for MuDoC 2.0 will focus on improving retrieval pipelines, leveraging cognitive learning theories for more effective multimodal responses, enhancing source attribution, extending summarization features to images and cross-page text, and providing user controls to manage cognitive load, ultimately aiming to solidify MuDoC's role as a reliable and engaging tool for enhanced interactions with course textbooks.

Acknowledgments

We are grateful for the support provided by National Science Foundation under Grant No. 2247790 and 2112532. We also thank Sandeep Kakar and Lingqing Wang for their feedback on the study design.

References

[1] K. Taneja, P. Maiti, S. Kakar, P. Guruprasad, S. Rao, A. K. Goel, Jill Watson: A Virtual Teaching Assistant Powered by ChatGPT, in: International Conference on Artificial Intelligence in Education, 2024, pp. 324–337. URL: https://link.springer.com/chapter/10.1007/978-3-031-64302-6_23.

[2] M. Wölfel, M. B. Shirzad, A. Reich, K. Anderer, Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust, Big Data and Cognitive Computing 8 (2024) 2. URL: https://www.mdpi.com/2504-2289/8/1/2.

[3] K. Taneja, A. K. Goel, MuDoC: An Interactive Multimodal Document-grounded Conversational AI System, in: AAAI Symposium Series 2025 Machine Learning and Knowledge Engineering (MAKE) for Trustworthy Multimodal and Generative AI, 2025. URL: http://arxiv.org/abs/2502.09843.

[4] K. Taneja, A. Singh, A. K. Goel, Towards a Multimodal Document-grounded Conversational AI System for Education, in: International Conference on Artificial Intelligence in Education, 2025. URL: http://arxiv.org/abs/2504.13884.

[5] A. Olson, P. Maiti, A. Goel, The Textbook of Tomorrow: Rethinking Course Material Interfacing in the Era of GPT, 2025. URL: http://arxiv.org/abs/2501.03618, arXiv:2501.03618 [cs].

[6] S. Feng, H. Wan, C. Gunasekara, S. Patel, S. Joshi, L. Lastras, doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8118–8128. URL: https://aclanthology.org/2020.emnlp-main.652/. doi:10.18653/v1/2020.emnlp-main.652.

[7] T. Lv, Y. Huang, J. Chen, Y. Zhao, Y. Jia, L. Cui, S. Ma, Y. Chang, S. Huang, W. Wang, L. Dong, W. Luo, S. Wu, G. Wang, C. Zhang, F. Wei, KOSMOS-2.5: A Multimodal Literate Model, 2024. URL: http://arxiv.org/abs/2309.11419. doi:10.48550/arXiv.2309.11419, arXiv:2309.11419 [cs].

[8] Y. Xu, T. Lv, L. Cui, G. Wang, Y. Lu, D. Florencio, C. Zhang, F. Wei, LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding, 2021. URL: http://arxiv.org/abs/2104.08836. doi:10.48550/arXiv.2104.08836, arXiv:2104.08836 [cs].

[9] J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, Y. Dan, C. Zhao, G. Xu, C. Li, J. Tian, Q. Qi, J. Zhang, F. Huang, mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding, 2023. URL: http://arxiv.org/abs/2307.02499. doi:10.48550/arXiv.2307.02499, arXiv:2307.02499 [cs].

[10] Y.-J. Tseng, Y.-H. Lin, G. Yadav, N. Bier, V. Aleven, Curio: An On-Demand Help-Seeking System on iTextbooks for Accelerating Research on Educational Recommendation Algorithms, in: AIED'23: Workshop on Intelligent Textbooks, 2023.

[11] D. Saindane, S. P. Prajapati, S. Das, Converting Physical Textbooks into Interactive and Immersive 'Phygital' Textbooks: A Proposed System Architecture Design for Textbook Companion Apps, in: AIED'23: Workshop on Intelligent Textbooks, 2023.

[12] R. E. Mayer, Multimedia learning, in: Psychology of Learning and Motivation, volume 41, Academic Press, 2002, pp. 85–139. URL: https://www.sciencedirect.com/science/article/pii/S0079742102800056.

[13] M. T. Thielsch, S. M. Meeßen, G. Hertel, Trust and distrust in information systems at the workplace, PeerJ 6 (2018) 5483. URL: https://peerj.com/articles/5483/.

[14] C. Ou, A. Goel, D. Joyner, Towards a Pedagogical Framework for Designing and Developing iTextbooks, in: AIED 2022: Workshop on Intelligent Textbooks, 2022.

[15] A. Goel, D. Joyner, Knowledge-based Artificial Intelligence: Cognitive Systems (e-book), Georgia Institute of Technology, 2018. URL: https://cs7637.netlify.app/Ebook/kbai_ebook-compressed.pdf.

[16] P. A. Kirschner, Stop propagating the learning styles myth, Computers & Education 106 (2017) 166–171. URL: https://www.sciencedirect.com/science/article/pii/S0360131516302482. doi:10.1016/j.compedu.2016.12.006.
[17] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. R. Naik, P. Cai, A. Gliozzo, Re2G: Retrieve, Rerank, Generate, in: NAACL 2022, 2022, pp. 2701–2715.
[18] E. Vendrow, O. Pantazis, A. Shepard, G. Brostow, K. E. Jones, O. Mac Aodha, S. Beery, G. Van Horn, INQUIRE: A Natural World Text-to-Image Retrieval Benchmark, Advances in Neural Information Processing Systems 37 (2024) 126500–126514. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/e4ad9c75f0d60ed75700f020adb3f705-Abstract-Datasets_and_Benchmarks_Track.html.
[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html?ref=https://githubhelp.com.
[20] T. B. Brown, B. Mann, N. Ryder, et al., Language Models are Few-Shot Learners, in: NeurIPS 2020, 2020.
[21] X. Ma, K. Sun, R. Pradeep, J. Lin, A Replication Study of Dense Passage Retriever, 2021. URL: http://arxiv.org/abs/2104.05740. doi:10.48550/arXiv.2104.05740, arXiv:2104.05740 [cs].
[22] J. Sweller, Cognitive Load Theory, in: Psychology of Learning and Motivation, volume 55, Academic Press, 2011, pp. 37–76. URL: https://www.sciencedirect.com/science/article/pii/B9780123876911000028.