EduMUSE: A Multimodal Educational Dataset with Automatically Extracted Instructional Context

Andreea Dutulescu and Stefan Ruseti and Mihai Dascalu
National University of Science and Technology POLITEHNICA Bucharest
{andreea.dutulescu, stefan.ruseti, mihai.dascalu}@upb.ro

Danielle S. McNamara
Arizona State University
danielle.mcnamara@asu.edu

Abstract

Research in AI applied to education increasingly relies on large-scale, high-quality datasets to support the development and evaluation of learning analytics and intelligent educational systems. Open educational resources provide a promising foundation, yet few datasets integrate structured instructional content with assessment materials in a multimodal form. In this study, we introduce a large-scale multimodal educational dataset (EduMUSE - Educational Multimodal Understanding & Solution Dataset) constructed from OpenStax undergraduate textbooks across multiple domains. The dataset integrates hierarchically structured instructional text, figures, exercises, and, when available, official solutions. For exercises with solutions, we introduce an automatic method that associates each exercise with a focused instructional subsection rather than entire textbook chapters, estimating subsection relevance via solution likelihood under candidate contexts using a vision–language model. We analyze the impact of contextualization on the behavior of vision–language models across different contexts. Results indicate that subsection-level instructional context has a measurable impact on model performance, with variation across model scales and task formulations. The dataset and code are released as open source at https://github.com/upb-nlp/BEA-EduMUSE/ to support reproducible research in multimodal educational modeling and to facilitate generating similar datasets using our approach.

1 Introduction

The effectiveness of educational technologies largely depends on the quality and structure of the datasets used for training and evaluation (Chen et al., 2025). This dependency has become more evident as the use of Large Language Models (LLMs) in educational applications, such as question answering, question generation, tutoring systems, and automated assessment, has increased (Wang et al., 2024; Dong et al., 2024; García-Méndez et al., 2025). LLMs typically require large amounts of data and are sensitive to noise, inconsistencies, and misalignment with instructional goals. As a result, the availability of clean, well-organized, and pedagogically relevant datasets remains an important factor in NLP research for education.

Many existing datasets for educational question answering and question generation are created using crowdsourcing platforms. While this approach facilitates scalable data collection, it often involves contributors without formal training in education or domain-specific expertise (Díaz et al., 2022; Klie et al., 2024). Moreover, the availability of datasets targeting college-level education is more limited, where materials often involve formal reasoning, domain-specific terminology, and structured explanations, which increases the difficulty of dataset construction and annotation.

In addition to textual content, many academic disciplines rely extensively on visual elements such as diagrams, figures, graphs, and illustrations (Jewitt, 2012). These elements play a central role in instruction and assessment across domains such as mathematics, physics, biology, chemistry, and engineering. However, a large portion of existing educational datasets remains text-centric, which constrains research on multimodal models that integrate visual and textual information in learning-oriented tasks.

To help address these gaps, we construct a large-scale multimodal educational dataset based on college-level textbooks from OpenStax (OpenStax, 2026). The dataset includes textbook text, images, exercises, and corresponding solutions across multiple domains. By sourcing content from openly licensed instructional materials, the dataset remains aligned with established curricula while enabling its reuse for research purposes. The inclusion of images supports the study of multimodal educational

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 358–367

tasks. An additional challenge in educational question answering is identifying the relevant instructional context required to solve a given exercise. In realistic learning scenarios, learners are expected to rely on specific portions of the instructional material rather than the full textbook. To model this aspect, we introduce an automatic, LLM-based method for extracting instructional context. The method evaluates candidate text segments based on the likelihood that an LLM produces a correct answer.

The main contributions of this work are as follows:

- The construction of EduMUSE, a large-scale, college-level multimodal educational dataset derived from OpenStax textbooks, including instructional text, images, exercises, and reference solutions across multiple domains
- An automatic, likelihood-based method for associating exercises with fine-grained instructional subsections, enabling compact contextual grounding for educational question answering
- An empirical analysis of vision–language model performance that quantifies the impact of instructional context, visual information, and model scale on solution likelihood and multiple-choice accuracy

2 Related Work

Large-scale question-answering datasets have played a central role in advancing machine reading comprehension and, more recently, LLMs. Early corpora primarily focused on factoid question answering over encyclopedic text, whereas more recent efforts have explored multi-hop reasoning, educational assessment, and synthetic data generation. However, existing datasets often trade off between scale, pedagogical structure, reasoning depth, and the explicit inclusion of instructional context and solutions.

Among the earliest and most influential datasets is SQuAD by Rajpurkar et al. (2016), which introduced large-scale span-based question answering over Wikipedia articles. Questions were designed by crowdworkers. Given its size and simplicity, SQuAD became a de facto benchmark for reading comprehension models. However, its questions largely emphasize local information retrieval and lexical matching, with limited requirements for complex reasoning or synthesis across sources. To address these limitations, Yang et al. (2018) introduced HotpotQA, which extended the single-document paradigm by introducing multi-hop question answering over Wikipedia articles. Questions require aggregating evidence across passages, and the dataset explicitly annotates supporting facts, enabling more fine-grained evaluation of reasoning. While HotpotQA represents a significant step toward compositional reasoning, it remains grounded in encyclopedic knowledge and inherits the limitations of crowd-authored questions, with less attention paid to pedagogical intent and question quality. Moreover, both SQuAD and HotpotQA are now relatively dated and were designed prior to the emergence of instruction-following LLMs, limiting their suitability for evaluating modern educational reasoning capabilities.

In contrast to crowdsourced encyclopedic datasets, Xu et al. (2022) adopted an explicitly educational perspective for FairytaleQA. The dataset is curated by educational experts and is based on children's storybooks, with questions designed to assess narrative understanding and basic comprehension skills appropriate for young learners. While FairytaleQA benefits from expert-driven design and pedagogical alignment, its scope is intentionally narrow: questions are simple, targeted at early education, and do not capture advanced academic domains.

A broader educational focus is found in EduQG by Hadifar et al. (2023), who introduced a multiple-choice question dataset derived from OpenStax textbooks. Questions are annotated with Bloom's taxonomy levels (Bloom et al., 1956), enabling analysis of cognitive complexity across formats. EduQG represents an important step toward structured educational QA, particularly through its explicit linkage to pedagogical theory. However, the dataset focuses exclusively on multiple-choice questions, limiting its applicability for tasks involving explanatory reasoning or open-ended answer generation.

In a recent benchmark, Dinh et al. (2024) targeted higher education and expert-level assessment. SCiEx targets university exam questions and covers domains such as Artificial Intelligence, Natural Language Processing, Human–Computer Interaction, and Databases. Despite the difficulty and topical diversity of the questions, the dataset notably lacks accompanying contextual material, requiring models to rely almost entirely on parametric knowl-

edge rather than contextualized reasoning.

Similarly, Singh et al. (2024) propose SCIDQA, a deep reading comprehension dataset derived from scientific papers, with questions originating from peer-review discussions and answers from author responses. While positioned as a dataset for deep scientific understanding, the alignment between questions and the provided paper content is not always explicit. In many cases, answering a question may require background knowledge or interpretative reasoning beyond what is directly stated in the text, raising questions about the dataset's suitability for controlled reading comprehension evaluation. As reasoning becomes more complex, Zhu et al. (2024) introduced FanOutQA, a multi-hop, multi-document QA benchmark in which answering a question requires aggregating information from a large number of articles. The dataset explicitly decomposes each question into intermediate sub-questions and associated documents, offering transparency into the reasoning process. Evaluation across several LLMs reveals substantial performance gaps, highlighting the difficulty of large-scale information aggregation. However, FanOutQA contains only development and test splits and is limited in size, making it unsuitable for model training and educational use. Additionally, the questions are primarily factual and analytic rather than pedagogically motivated.

An alternative line of work involves generating synthetic datasets. Gong et al. (2025) proposed PhantomWiki, a framework for on-demand dataset creation by generating a synthetic universe of entities, relations, and documents, followed by grammatically generated questions. This approach enables precise control over the complexity of reasoning and the difficulty of retrieval. However, the resulting texts and questions are highly artificial, with limited linguistic naturalness and no grounding in authentic educational content. As a result, while PhantomWiki is valuable for the evaluation of controlled reasoning, it is less suitable for modeling real-world learning scenarios. Similarly, Naeem et al. (2025) introduced a synthetic educational QA dataset to assess grade-level adaptability in LLMs. While it addresses an important dimension of educational alignment, the reliance on synthetic generation raises concerns regarding linguistic diversity, authenticity, and alignment with real instructional materials.

Taken together, existing educational datasets span a wide spectrum of scale, reasoning depth, and pedagogical grounding. Large encyclopedic datasets offer scale but limited educational structure; curated educational datasets provide pedagogical alignment but are often restricted in complexity or format; expert-level and scientific benchmarks lack explicit context; and synthetic datasets enable control at the cost of ecological validity. These limitations motivate the need for a multimodal educational dataset that jointly provides rich context, well-designed questions, and explicit solutions, enabling the study of grounded reasoning, explanation generation, and learning-oriented evaluation in realistic educational settings.

3 EduMUSE Construction

This section describes the construction of the multimodal educational dataset - EduMUSE - Educational Multimodal Understanding & Solution Dataset) - and the procedure for augmenting exercises with relevant instructional context. The method consists of two main components: large-scale scraping and normalization of college-level instructional materials from OpenStax textbooks, and an automatic context selection procedure that identifies the most relevant instructional subsections for each exercise using a likelihood-based criterion derived from a vision-language model.

3.1 Dataset Scraping

3.1.1 Source Selection

We construct EduMUSE using textbooks published by OpenStax, an initiative that provides openly licensed, college-level textbooks across multiple academic disciplines (OpenStax, 2026). OpenStax materials are aligned with standard undergraduate curricula and include a rich combination of instructional text, figures, worked examples, exercises, and solutions, making them well-suited for educational NLP tasks. EduMUSE covers textbooks from the following domains: Business, Computer Science, Humanities, Mathematics, Nursing, Science, and Social Sciences. These domains were selected to ensure diversity in reasoning styles, representational formats, and instructional practices, ranging from formal symbolic reasoning to conceptual and narrative explanations.

3.1.2 Content Extraction

OpenStax does not provide a dedicated API or bulk export mechanism for extracting structured datasets. As a result, data collection was performed

by directly scraping the textbooks in HTML representations. Although the absence of an official scraping interface increases implementation complexity, the HTML pages follow a consistent structural schema, enabling reliable content extraction. Specifically, the HTML markup includes identifiable tags and class attributes corresponding to chapters, sections, subsections, paragraphs, figures, exercises, and solutions. These structural cues were leveraged to systematically extract and hierarchically organize the textbook content.

For each textbook, we extracted the instructional content at the paragraph level, preserving the original hierarchical organization of chapters, sections, and subsections. Paragraphs were cleaned by removing extraneous HTML elements, navigation artifacts, and formatting tags, while retaining inline mathematical expressions and emphasized text. Figures and images embedded in the instructional content were also extracted. For each figure, we retained both the image itself and its associated alternative (alt) text and caption. The alt text in OpenStax textbooks typically provides a detailed description of the visual content. Retaining this serves two purposes: it supports multimodal learning scenarios for models that process images, and it provides a textual proxy for visual information when evaluating text-only models.

3.1.3 Exercise and Solution Extraction

In addition to instructional content, we extracted end-of-chapter exercises. Exercises were identified using HTML tags specific to assessment content and are associated with a corresponding chapter. Each exercise was cleaned to remove residual HTML markup, footnotes, and styling elements. Mathematical expressions originally encoded using MathJax were converted into a normalized textual representation to ensure compatibility with language models. This normalization step produces equation strings that are syntactically explicit and easily parsable by LLMs. When available, we also scraped the official solutions linked to the exercises. Exercises with solutions typically include a hyperlink to a separate solution page, which was followed and processed in the same manner as the exercise text. Solution content was cleaned using the same normalization pipeline applied to exercises. For exercises that included images or graphs in the problem statement, the visual elements and their corresponding alt text were retained, as they may be necessary to solve the exercise.

3.1.4 Limitations of Chapter-Level Context

While exercises are associated with chapters in the textbook structure, chapters are often extensive, containing thousands of tokens. Providing the full chapter as context for an exercise is impractical for many language models due to context length limitations. Even for long-context models, such large contexts can lead to diluted attention, increased inference cost, and spurious reliance on irrelevant information. This motivates the need for a more granular approach to identifying the instructional content most relevant to each exercise.

3.2 Context Augmentation

3.2.1 Motivation

In realistic educational settings, learners solve exercises by relying on specific instructional segments rather than the entire textbook chapter. However, existing datasets typically associate exercises with coarse-grained contexts (e.g., entire chapters or no context). To model a more realistic and pedagogically meaningful scenario, we aim to identify the most relevant instructional subsection for each exercise.

Each textbook chapter is hierarchically organized into sections, which are further subdivided into subsections. These subsections generally correspond to coherent instructional units focused on specific concepts, methods, or examples. As such, we consider subsections as candidate instructional contexts and aim to select, for each exercise, the subsection that provides the most relevant information for solving it.

3.2.2 Relevant Subsection Selection

The context augmentation procedure is applied to exercises for which an official solution is available. The presence of a reference solution enables an objective evaluation of how well a given instructional context supports a model's correct answer generation.

For a given exercise with question text e, a reference solution s, and a set of candidate subsections $c_1, ..., c_N$ from the corresponding chapter, we evaluate the relevance of each subsection by measuring how well it supports the generation of the correct solution. Formally, for each candidate subsection c_i, we compute the conditional probability: $P(s \mid c_i, e)$, where c_i and e are provided in the prompt, as instructional context and exercise statement. This probability is estimated using the

token-level logits of an instruction-tuned vision-language model. The probability is computed by decomposing the solution into a sequence of tokens $s = (s_1, ..., s_T)$ and summing the log-probabilities assigned by the model:

$$\log P(s \mid c_i, e) = \sum_{t=1}^{T} \log P(s_t \mid s_{<t}, c_i, e) \quad (1)$$

where e is the target exercise, c_i is the instructional context, s is the annotated solution, and t iterates over the solution's tokens.

This formulation captures how consistently the model predicts the annotated solution when conditioned on a given subsection.

For each exercise, we compute $P(s \mid c_i, e)$ for all candidate subsections within the chapter. The subsection that yields the highest probability is selected as the most relevant instructional context: $c^* = \arg\max_{c_i} P(s \mid c_i, e)$. This selected subsection is then associated with the exercise as its contextual grounding.

3.2.3 Experimental Setup

We employ Qwen2-VL-7B-Instruct[1] (Team et al., 2024), an instruction-tuned vision-language model that supports both textual and visual inputs. This choice allows the context evaluation procedure to incorporate not only textual instructional content but also associated images when present. Moreover, Qwen models allow for multiple images in the input, unlike other VLM families. For exercises or subsections that include figures, the images are provided according to the model's input specifications and are correctly positioned in their original locations within the text.

3.2.4 Outcome

After applying the previous procedure, each exercise with an available solution in EduMUSE becomes paired with a specific instructional subsection that is empirically validated to be maximally supportive of correct answer generation by a vision-language model. This context augmentation enables more realistic evaluation of educational question answering and reasoning tasks while reducing context length and mitigating the effects of irrelevant instructional material.

[1] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

3.3 Dataset Characterization

This subsection provides a quantitative overview of EduMUSE after scraping, normalization, and context augmentation. EduMUSE was constructed from 39 textbooks across multiple academic domains. From these sources, we extracted 40,971 exercises in total. Among these, 7,195 exercises (17%) are multiple-choice questions (MCQs), reflecting the diversity of assessment formats present in the original textbooks.

A subset of the exercises includes official reference solutions. In total, 15,297 exercises (37%) are paired with solutions, of which 2,037 are MCQs. These exercises constitute the primary subset used for context augmentation, as a reference solution is required to compute solution likelihoods during subsection selection.

EduMUSE also contains substantial visual information. A total of 2,239 exercises include one or more images directly in the problem statement, such as diagrams, plots, or illustrative figures. In addition, 8,595 exercises are associated with instructional subsections that contain images. This enables the evaluation of multimodal reasoning scenarios in which visual content appears either in the question, the instructional context, or both.

Figure 1 illustrates the distribution of the number of exercises across the textbooks included in EduMUSE. The majority of textbooks contain a substantial number of exercises with official solutions, while a non-negligible subset includes no exercises with solutions. In addition, only a limited number of textbooks provide exercises in the form of multiple-choice questions. By contrast, a significant proportion of textbooks feature multiple exercises that incorporate visual elements, such as diagrams or figures, into the problem statement or instructional context.

In terms of instructional content granularity, textbook chapters are relatively long, averaging 13,133 words and reaching a maximum of 64,511 words. By contrast, subsections are significantly shorter, with an average length of 200 words and a maximum of 5,921 words. This disparity highlights the practical importance of subsection-level context selection, as subsections provide focused instructional units that are substantially more manageable for language and vision-language models than full chapters.

Overall, EduMUSE supports a wide range of educational modeling scenarios, including text-only

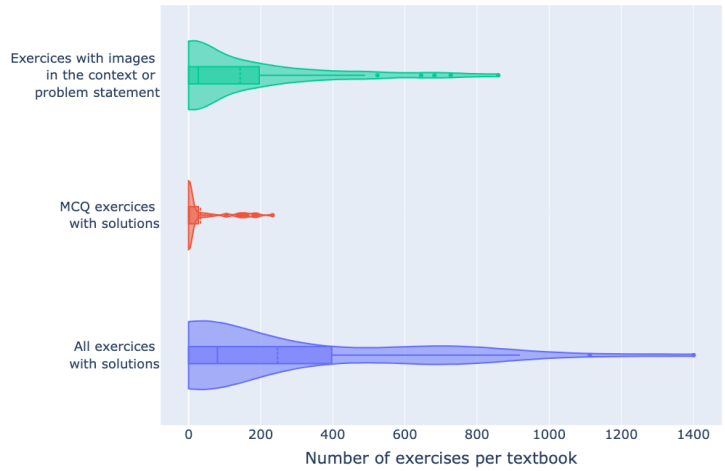

Figure 1: The distribution of the number of exercises per textbook in the OpenStax dataset.

and multimodal question answering, solution generation, and context-aware reasoning. The combination of diverse domains, varied exercise formats, explicit solution annotations, and fine-grained instructional context makes EduMUSE suitable for evaluating models under realistic instructional conditions.

4 EduMUSE Validation via VLM Probes

We use vision–language models (VLMs) on EduMUSE as probes to validate that it encodes meaningful instructional structure, with particular focus on the contributions of visual information and the selected instructional context (Section 3.2). We employ a VLM of a different family than the language model used for context augmentation while constructing EduMUSE. This separation ensures that the instructional signals encoded in the dataset generalize to newer and stronger vision–language models rather than reflecting model-specific biases. Specifically, we use Qwen3-VL-[8/32]B-Instruct [2] (Yang et al., 2025) as the answering model, evaluating two parameter sizes: 8B and 32B.

4.1 Probabilistic Evaluation of Annotated Solutions

In the first evaluation setup, we consider exercises for which a reference solution is explicitly annotated in the dataset. For each such exercise, we compute the probability that the model generates the correct solution, following Equation 1 described in Section 3.2. This formulation allows us to assess how different forms of context influence the model's likelihood of producing the correct answer, without relying on discrete generation or heuristic matching.

We consider three contextual conditions: a) *full* multimodal context: the complete instructional context, including both text and images; b) t*ext-only* context: the full textual context, with images removed but their associated alternative text retained; and c) *no* context: the exercise prompt alone, without any instructional context. For each exercise and model size, we compute the token-level likelihood that the model assigns to the reference (annotated) solution, given the exercise prompt and a specified context. This yields three probabilities: one with the full multimodal instructional context, one with text-only context, and one with no additional context. To enable meaningful comparison across exercises with widely varying base difficulty, we compute within-exercise probability ratios between (i) the full multimodal context and the text-only context, and (ii) the full multimodal context and the no-context condition.

To summarize these ratios across EduMUSE, we report the median (50th percentile) ratio for each model and comparison. Using the median reduces sensitivity to extreme values arising from individual exercises where probabilities are unusually small or large. Therefore, the resulting values represent a *typical multiplicative increase in solution likelihood attributable to instructional and/or visual context*. Table 1 reports these median ratios for both the 8B and 32B models.

The purpose of this evaluation is not to achieve

[2] https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct, https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct

state-of-the-art problem-solving accuracy, but to verify that EduMUSE encodes instructional structure in a form that is operationally detectable by contemporary multimodal models and that confidence substantially increases when appropriate context is provided. If the extracted instructional context is meaningful, then conditioning on it should systematically increase the likelihood assigned to expert-annotated solutions—particularly in exercises where visual material plays a pedagogical role. The observed probability gains therefore serve as evidence that EduMUSE captures instructional signals that models can exploit, rather than as a metric of absolute task performance.

The results indicate a consistent and substantial benefit from providing instructional context, particularly when visual information is included. For the larger Qwen3-32B model, the median probability of generating the correct solution is approximately 13× higher when the full multimodal context is provided compared to a text-only context without images. In contrast, the smaller Qwen3-8B model exhibits a much stronger dependency on visual information, with a 366× increase in median probability in the same comparison. A similar trend is observed when comparing the full multimodal context to the no-context condition, with probability ratios of 94× for the 32B model and 462× for the 8B model.

These results suggest that larger models are more robust to reductions in available instructional content. The Qwen3-32B model likely compensates for missing contextual information by leveraging external knowledge encoded in its parameters, resulting in smaller performance degradation when images or context are removed. Conversely, the smaller Qwen3-8B model relies more heavily on explicit instructional signals, particularly multimodal cues, to produce correct solutions. Nevertheless, across both model scales, the inclusion of multimodal instructional content consistently yields higher probabilities of correct answers. This highlights the importance of multimodal context for educational question answering across diverse exercise types, including structured problems and open-ended responses.

4.2 Multiple-Choice Question Answering Performance

In the second evaluation setup, we assess model performance on multiple-choice questions (MCQs). For each MCQ, models are prompted to select the correct answer option, constrained to output only a single letter corresponding to one of the choices. To ensure consistent and unambiguous outputs, we employ structured decoding that restricts generation to the available answer set.

As in the probabilistic setup, we consider three contextual conditions: full multimodal context, text-only context (without images but including alternative text), and no context. Performance is measured by accuracy, defined as the percentage of exercises answered correctly. In this setting, multiple-choice answering is used as a stress test for the dataset's contextual structure under constrained output conditions. Unlike probabilistic solution likelihood, this formulation requires models to compress potentially rich instructional information into a single discrete decision, making it particularly sensitive to context length and model capacity.

Table 2 reports accuracy results across all MCQ exercises in the dataset, while Table 3 considers only exercises whose instructional context contains at least one image. This stratification enables a more direct analysis of the impact of visual information on question answering performance. Comparisons across model sizes and context conditions reveal how scaling and multimodal inputs jointly affect VLMs' capability to solve college-level textbook problems.

In contrast to the probabilistic evaluation presented earlier, this setup requires explicit answer generation under a constrained output format. Under these conditions, we observe that the smaller Qwen3-8B model is negatively affected by the inclusion of large instructional contexts. For both the full set of exercises and the subset containing images, the 8B model achieves the highest accuracy in the no-context setting, with performance decreasing as more context is provided. This suggests that, for smaller models, long or information-dense contexts may negatively impact the generation. Nevertheless, this behavior does not impact subsection selection, as the selection is performed based on likelihood derived from model logits rather than on the generated output itself.

This behavior is consistent with findings reported by Hsieh et al. (2024) who show that smaller models experience more pronounced performance degradation as context length increases. Limited parameter capacity and narrower effective attention spans reduce their capability to attend to relevant information in long contexts. As a result, additional

Context Comparison	Qwen3-8B	Qwen3-32B
Full vs. Text-only Context	366	13
Full vs. No Context	462	94

Table 1: Median probability ratios (50th percentile) comparing full multimodal context against reduced-context settings for Qwen3 models.

Context Condition	Qwen3-8B	Qwen3-32B
Full Multimodal Context	31.66%	41.04%
No Context	37.65%	40.89%

Table 2: Multiple-choice question accuracy for Qwen3 models under different instructional context conditions across all exercises.

Context Condition	Qwen3-8B	Qwen3-32B
Full Multimodal Context	45.05%	58.97%
Text-only Context	47.39%	58.89%
No Context	53.74%	58.81%

Table 3: Multiple-choice question accuracy for Qwen3 models on exercises whose instructional context contains at least one image.

instructional content, while potentially useful, can hinder decision-making in constrained generation tasks such as multiple-choice answering.

In contrast, the Qwen3-32B model exhibits stable or slightly improved performance when an instructional context is provided. For the full exercise set, accuracy is comparable across the whole-context and no-context conditions, whereas for exercises containing images, the full multimodal context yields slightly higher accuracy. This indicates that larger models are better able to leverage extended and multimodal contexts, likely due to increased parameter capacity and more expressive attention mechanisms that enable effective filtering of relevant information.

Finally, across all evaluation conditions, we observe a clear performance gain when moving from the 8B to the 32B model, confirming the expected scaling effects of model size on educational question answering accuracy.

5 Conclusions and Future Directions

This work introduces EduMUSE, a large-scale, multimodal educational dataset derived from college-level OpenStax textbooks and designed to support research on solution generation and context-aware question answering in realistic instructional settings. By systematically extracting textbook text, images, exercises, and official solutions across multiple academic domains, the dataset addresses several limitations of existing educational benchmarks, including the lack of pedagogical grounding, insufficient multimodal content, and coarse or absent instructional context.

A key contribution of this work is the proposed automatic context augmentation method, which associates each exercise with a focused instructional subsection selected based on the solution likelihood predicted by a vision–language model. This approach produces compact contexts that better reflect how learners engage with educational materials. The resulting subsection-level grounding mitigates the challenges posed by long chapters while preserving pedagogical coherence. By validating the dataset through controlled model probes, this work aligns with goals of educational technology: understanding how structure, context, and representation shape performance in learning-oriented tasks.

Overall, the dataset and accompanying analyses provide a foundation for studying multimodal educational reasoning under controlled yet realistic conditions. The combination of authentic instructional content, explicit solutions, fine-grained context, and diverse domains enables future work on curriculum-aware modeling, explanation generation, context selection, and the interaction between visual and textual information in learning-oriented tasks. The dataset also offers opportunities for investigating the impact of context length and model scaling behavior.

Limitations

Despite the strengths of EduMUSE, several limitations should be acknowledged.

First, although the dataset spans multiple academic domains, it is constructed from a finite set of OpenStax textbooks that primarily cover widely taught undergraduate subjects. As a result, the dataset may not fully capture the breadth of specialized, advanced, or less commonly taught topics. Extending the dataset to include a broader range of disciplines, including graduate-level materials and niche subject areas, would improve its coverage and applicability.

Second, the proposed context selection method relies on model-derived likelihoods to identify the most relevant instructional subsection for each exercise. While this approach provides a principled and scalable mechanism for automatic context assignment, it is inherently dependent on the behavior and calibration of the underlying model. In particular, although likelihood-based scoring using logits offers a reasonable proxy for contextual relevance, it does not disentangle whether improvements arise from genuinely informative context or from increased model confidence in generating specific token sequences.

Overall, these limitations highlight important directions for future work. Addressing these aspects should aim to improve the reliability and generalizability of multimodal educational datasets and understanding on how models leverage instructional context in realistic learning scenarios.

Acknowledgments

The research reported here was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University, and the grant of the Academy of Romanian Scientists, AOSR-TEAMS-IV Edition 2025-2026 "Digital Transformation in Science". The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay, New York.

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. 2025. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Remi Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351.

Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, and 1 others. 2024. Sciex: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610.

Bingyu Dong, Jie Bai, Tao Xu, and Yun Zhou. 2024. Large language models in education: A systematic review. In *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 131–134. IEEE.

Silvia García-Méndez, Francisco de Arriba-Pérez, and María del Carmen Somoza-López. 2025. A review on the use of large language models as virtual tutors. *Science & Education*, 34(2):877–892.

Albert Gong, Kamilė Stankevičiūtė, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P Gomes, and Kilian Q Weinberger. 2025. Phantomwiki: On-demand datasets for reasoning and retrieval evaluation. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Eduqg: A multi-format multiple-choice dataset for the educational domain. *Ieee Access*, 11:20885–20896.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *CoRR*.

Carey Jewitt. 2012. *Technology, literacy, learning: A multimodal approach*. Routledge.

Jan-Christoph Klie, Richard Eckart De Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Numaan Naeem, Abdellah El Mekki, and Muhammad Abdul-Mageed. 2025. Eduadapt: A question answer benchmark dataset for evaluating grade-level adaptability in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34224–34251.

OpenStax. 2026. Openstax: Free and open college textbooks. https://openstax.org/.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. Scidqa: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, and 1 others. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37.