Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

LLM-enhanced Math Text Extracting and Keyword-based Hierarchical Labeling for Digital Learning Infrastructure

Taisei YAMAUCHI[a*], Brendan FLANAGAN[b, c], & Hiroaki OGATA[c]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Institute for Liberal Arts and Sciences, Kyoto University, Japan*
[c]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*yamauchi.taisei.28w@st.kyoto-u.ac.jp

Abstract: As education becomes increasingly digitized, organizing digital learning materials into curriculum-based units is essential but often burdensome for educators. This study proposes an automatic method for labeling mathematics problems in PDF format using large language models (LLMs) and curriculum keywords. Using OpenAI's o4-mini, we extracted text from MEXT-approved junior high school textbooks and exercise books with high accuracy (98.9% and 99.7%). Unit labels were then assigned by combining keyword-based filtering with embedding similarity (text-embedding-3-small). Compared with a baseline without keyword filtering, expert evaluation favored the keyword-based method (183 vs. 132 cases), confirming that keywords enhance classification accuracy. These results demonstrate that LLM-based extraction is practical for classroom use, requiring only minor manual corrections, and that unit-specific vocabulary contributes to accurate hierarchical labeling. Future work will extend this framework toward content management of LLM-generated materials and unified log analysis to promote personalized learning pathways.

Keywords: Automatic unit labeling, large language models, mathematics learning content, personalized learning environments

1. Introduction

Within digital learning environments, there is an emerging focus on analyzing learning behaviors by labeling educational content with unit information. This approach is particularly valuable in mathematics education (Hussein, 2023), as unit labeling supports both learners and educators by enabling systems to provide tailored assistance based on content and learner data (Vovides et al., 2007). Previous studies have shown that associating unit labels with PDF learning content can help reveal patterns in learning activities (Wang et al., 2022). In parallel, research in knowledge tracing has demonstrated that regression and classification models can estimate learners' knowledge levels from sparse log data, effectively incorporating multiple knowledge elements and skill dimensions (Vie & Kashima, 2019; Wang et al., 2023).

Further, visualizing the relationships among topics can enhance the consistency of assessments and provide meaningful feedback to learners (Khosravi & Cooper, 2018). The classification of exercises has also been shown to be beneficial in recommending tasks that strengthen conceptual understanding, with evidence suggesting that students prefer recommendations that include explanatory components (Yamauchi et al., 2024). Other work has emphasized the value of unit-tree structures in making personalized education systems more transparent (Sosnovsky & Brusilovsky, 2015), as well as the importance of recommendation systems that support students in selecting suitable content based on both difficulty and individual preferences (Christudas et al., 2018). Research has also explored the extraction of units from learning materials to construct knowledge-structure representations, which can increase learners' awareness of their educational trajectory (Flanagan et al., 2019).

To conduct advanced analyses of student and teacher logs, learning materials must be classified into curriculum knowledge units. However, such unit information is often absent

from the content itself, and manual annotation imposes a heavy workload on educators (Flanagan et al., 2023; Schubotz et al., 2020). This has driven growing demand for automatic labeling methods. While automatic unit labeling has been shown to reduce the burden on experts (Schubotz et al., 2020), further advances are required to improve accuracy in order to make such systems practical.

Although prior research demonstrated that mathematical formulas can be reliably processed from well-structured PDF files (Date & Isozaki, 2015), challenges remain when dealing with the diverse formats of user-uploaded PDFs. In these cases, information extraction often yields incomplete results (Abekawa & Aizawa, 2016), which complicates the implementation of automatic unit-labeling systems in educational contexts. To address this, Yamauchi et al. (2023) proposed an n-gram–based labeling method that achieved strong performance, even surpassing embedding models, for short and incomplete texts. In addition, Yamauchi et al. (2025) showed that perceptron models with bigram features performed well across diverse mathematics learning materials, demonstrating the potential of lightweight text-based methods for automatic unit labeling.

Recent advances in large language models (LLMs) have significantly improved the accuracy of text extraction. Alongside this progress, embedding techniques that capture contextual meaning, beyond strict string matching, have become feasible for analyzing educational texts. There is also research showing that using keywords improves labeling accuracy (Flanagan et al., 2023), but there is room for verification as to whether this is effective using LLM technology. In this study, we propose a method that leverages LLMs for extracting mathematical text from PDF exercise materials and applies keyword-based unit labeling to systematically organize the extracted problems.

RQ1: Can LLM-based methods accurately extract text from diverse mathematics materials?
RQ2: Does combining keyword information with LLM embeddings improve labeling accuracy compared to embeddings alone?

2. Method and Result

2.1 Task description

The task in this study is to label mathematics problems stored in PDF format with a hierarchical structure consisting of a main unit, sub unit, and subsub unit. This involves two main steps: (1) extracting mathematical text from the PDF files, and (2) comparing the extracted text with the corresponding content in the reference textbook to assign the appropriate hierarchical labels.

2.2 Data Preparation

This study utilized three types of resources: Japanese junior high school (7th–9th grade) mathematics textbooks, exercise books, and a keywords list. The textbook used in this study was a Ministry of Education, Culture, Sports, Science and Technology (MEXT) approved Japanese mathematics textbook. The PDF version of the textbook contained 625 pages and was organized into a hierarchical structure consisting of 22 *main units*, 55 *sub units*, and 109 *subsub units*. The exercise book was also based on a MEXT-approved textbook; however, it was not the same textbook as the one used in this study. This choice was motivated by the fact that some schools adopt a combination in which the exercise book corresponds to a different MEXT-approved textbook than the main textbook used in class. Due to this difference, the hierarchical structure of the units in the exercise book does not perfectly align with that of the main textbook. The PDF version of the exercise book consisted of 1,769 pages. The keywords list was derived from the “Multilingual Junior High School Mathematics Glossary” (https://www.mext.go.jp/a_menu/shotou/clarinet/003/001/011/003.htm) published by MEXT. This glossary comprehensively covers all terminology used in junior high school mathematics. A total of 738 keywords were obtained from the glossary.

2.3 Extraction of mathematical text

To extract text from the mathematics textbooks and exercise books, we employed a LLM based prompting approach. Specifically, we used o4-mini, a member of OpenAI's o-series models. The o-series models are capable of reasoning and can perform a wide range of tasks with relatively high accuracy; however, among these, only the o4-series supports image input. Considering cost efficiency, o4-mini was selected for this study.

The text extraction algorithm is summarized in Table 1. First, each page of the PDF files was converted into high-resolution (450 dpi) JPEG images. For use as input to the LLM, these images were then converted into Base64-encoded strings and fed into the model.

Because both the textbooks and exercise books often contain diagrams, we designed separate prompts for extracting the main text and for extracting the images, as outlined in Table 2. The prompts were refined iteratively based on insights gained during preliminary extraction trials. For example, since fractional expressions were often extracted inaccurately, specific instructions on how to extract fractions were added to the prompts. Additionally, the model frequently returned "This page cannot be extracted" without performing extraction. By modifying the prompt to include the instruction "If extraction is not possible, please add the reason," the model unexpectedly began to produce extractions rather than refusal messages, which is a notable and interesting behavioral change. This final version of the prompt was adopted in our extraction pipeline.

Across the dataset, text was successfully extracted from 618 of 625 textbook pages (98.9%) and 1,763 of 1,769 exercise-book pages (99.7%).

Table 1. *Algorithm of extracting math text from PDF file*

	Algorithm
1	Convert each PDF to images (JPEG) at 450 dpi and save them in the image_file folder.
2	Process the JPEG files in image_file in name order:
3	Base64-encode the image.
4	Run the prompt 1 (in Table 2) to extract text and record both the raw result and a formatted version.
5	Run the prompt 2 (in Table 2) to extract figure information and record both the raw result and a formatted version.

Table 2. *Prompt for extraction of math text and image*

	Prompt
1	Your task is to extract the text contained in the image and output it as text. The following image contains a math problem and its solution. Please output the text that appears in this image. - For the solution process, output exactly the characters written in the image. - Write mathematical expressions in MathJax format, using the notation conventions used in MathJax for mathematical symbols. - Write fractions explicitly. For example, write `\frac{{1}}{{4}}` instead of `\frac14`, and write `\frac{{1}}{{2}}x^2` instead of `\frac12x^2`. - No markdown is needed. - If you cannot extract all of the characters, output only the characters you were able to extract. - If you cannot output the text contained in the image at all, explain why it cannot be done.
2	Your task is to output a textual description of the figure contained in the image. The following image contains a math problem and its solution. Please output the text that appears in this image. - You must provide an appropriate and uniquely determined description. - Output all information necessary to reproduce the figure, such as side lengths, coordinates, etc. - Write mathematical expressions in MathJax format. - Write fractions explicitly. For example, write `\frac{{1}}{{4}}` instead of `\frac14`, and write `\frac{{1}}{{2}}x^2` instead of `\frac12x^2`.

- No markdown is needed.
- If you cannot provide an output, explain why it cannot be done.
- Do not transcribe sentences; describe only the figure.

2.4 Calculation of keyword tf-idf

From the textbook text data, all keywords contained in the glossary were extracted by exact string matching. For each main unit, the corresponding block of text was then represented using a TF–IDF vectorizer, which generated vectorized representations with respect to the occurrence of each keyword.
As a result, 453 keywords appeared in at least one text segment. Among them, 158 keywords were unique to a single main unit. This finding indicates that the mathematics teaching materials contain multiple unit-specific terms, suggesting that certain vocabulary items are strongly characteristic of individual main units.

2.5 Extraction of mathematical text

The overall algorithm is summarized in Table 3. First, candidate main units were identified by matching the extracted text with the predefined keywords. Subsequently, sub units and subsub units were determined by calculating the cosine similarity between the text of each textbook page and the corresponding unit texts.

For vectorization of each page, we employed OpenAI's text-embedding-3-small model via the API. This embedding method was chosen because it enables fast transformation of text into vector representations while capturing not only lexical overlap but also contextual similarity, thereby improving the accuracy of unit assignment.

Table 3. *Algorithm of unit labeling to math texts*

	Algorithm
1	# Step 1: Main unit candidates from keywords
2	if exercise has keywords:
3	match keywords to TF-IDF matrix
4	score each main unit
5	max_score <- max score among all main units
6	candidates <- units with score > 0.75 × max_score
7	# Step 2: Main unit selection by similarity
8	reference_data <- reference_data with main_unit in candidates
9	compare exercise vector with reference_data (cosine similarity)
10	group similarities by main_unit
11	take median of top 3 similarities for each unit
12	max_median <- max median among all main units
13	estimated_units <- units with median > 0.95 × max_median
14	# Step 3: Subunit selection for each main unit
15	for each main unit in estimated_units:
16	group similarities by sub_unit
17	take median of top 3 similarities for each unit
18	max_median <- max median among all main units
19	estimated_subunits <- subunits with median > 0.95 × max_median
20	# Step 4: Subsubunit selection for each subunit
21	for each main unit in estimated_subunits:
22	group similarities by subsub_unit
23	max_median <- max median among all main units
24	estimated_subsubunit <- one subsubunit with max_median
25	estimeted_subunit <- subunit with estimated_subsubunit
26	estimeted_unit <- unit with estimated_subunit
27	return estimated_unit, estimated_subunit, estimated_subsubunit

2.6 Comparison and Evaliation

To evaluate the contribution of keywords to unit assignment accuracy, we compared the algorithm described in Table 3 with a variant that omitted Step 1, i.e., the narrowing of candidate main units by keyword matching. Between the two methods, differences in assigned units were observed for 315 out of 1,769 exercise-book pages. These pages were then annotated by a mathematics education expert, who determined which assignment was more appropriate for each case. The results showed that the keyword-based method was judged more appropriate in 183 cases, while the non-keyword method was preferred in 132 cases. This outcome indicates that incorporating keywords into the algorithm improves the accuracy of unit assignment.

3. Discussion and Limitations

Although prior research demonstrated that mathematical formulas can be reliably processed from well-structured PDF files (Date & Isozaki, 2015), challenges remain when dealing with the diverse formats of user-uploaded PDFs. In such cases, information extraction often produces incomplete results (Abekawa & Aizawa, 2016), which complicates the implementation of automatic unit-labeling systems in educational contexts. Moreover, to conduct advanced analyses of student and teacher logs, learning materials must be classified into curriculum knowledge units. However, unit information is often not embedded within the content itself, and manual annotation imposes a substantial workload on educators (Flanagan et al., 2023; Schubotz et al., 2020).

In this study, LLM-based methods achieved high extraction accuracy: 618 of 625 textbook pages (98.9%) and 1,763 of 1,769 exercise-book pages (99.7%) were successfully processed. Since the materials analyzed are those actually used in schools, this result implies that teachers would only need to manually revise approximately ten pages among thousands, making the system sufficiently practical for classroom use. With regard to labeling, incorporating keywords into the embedding-based approach yielded higher accuracy than embeddings alone. This improvement can be attributed to the presence of unit-specific keywords, which effectively contributed to the identification of main units.

Nevertheless, certain limitations must be acknowledged. Annotation was conducted by a single domain expert, and cases where both models produced the same prediction were not independently verified for correctness. These factors highlight important directions for future work, including multi-expert validation and broader evaluation of unit-labeling reliability.

4. Conclusion

This study proposes an automatic method for labeling mathematics problems in PDF format using LLMs and curriculum-based keywords. Applying o4-mini, we extracted text from junior high school textbooks and exercise books with high accuracy (98.9% and 99.7%). Unit labels were then assigned using keyword-based filtering and embedding similarity, with expert evaluation showing that incorporating keywords improved accuracy.

In the future, the system of this research can develop a content-management layer that curates, versions, and audits LLM-generated learning materials (Xing et al., 2025). In parallel, it can also unify learning logs across heterogeneous resources into a single, unit-aware record, enabling models to predict unit-level knowledge gaps for each learner and to drive more effective personalized sequencing, recommendations, and formative assessment (Takii et al., 2025).

Acknowledgements

This work was partly supported by JSPS JP25KJ1627, JP23K25698 and JP23H01001, JP23H00505, and CSTI SIP Program JPJ012347.

References

Abekawa, T., & Aizawa, A. (2016). SideNoter: Scholarly paper browsing system based on PDF restructuring and text annotation. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, 136–140.

Christudas, B. C. L., Kirubakaran, E., & Thangaiah, P. R. J. (2018). An evolutionary approach for personalization of content delivery in e-learning systems based on learner behavior forcing compatibility of learning materials. *Telematics and Informatics, 35*(3), 520–533. https://doi.org/10.1016/j.tele.2017.02.004

Date, I., & Isozaki, H. (2015). Detection of mathematical formula regions in images of scientific papers by using deep learning and OCR. *IEICE Technical Report, 2015*(4), 1–6. https://doi.org/10.1109/ACCESS.2019.2945825

Flanagan, B., Majumdar, R., Akçapınar, G., Wang, J., & Ogata, H. (2019). Knowledge map creation for modeling learning behaviors in digital learning environments. *Proceedings of LAK19: 9th International Conference on Learning Analytics and Knowledge*, 428–436.

Flanagan, B., Tian, Z., Yamauchi, T., Dai, Y., & Ogata, H. (2023). A human-in-the-loop system for labeling knowledge components in Japanese mathematics exercises. *Research and Practice in Technology Enhanced Learning, 19*(28). https://doi.org/10.58459/rptel.2024.19028

Hussein, H. B. (2023). Global trends in mathematics education research. *International Journal of Research in Educational Sciences, 6*(2), 309–319. https://doi.org/10.29009/ijres.6.2.9

Khosravi, H., & Cooper, K. (2018). Topic dependency models: Graph-based visual analytics for communicating assessment data. *Journal of Learning Analytics, 5*(3), 136–153. https://doi.org/10.18608/jla.2018.53.9

Sosnovsky, S., & Brusilovsky, P. (2015). Evaluation of topic-based adaptation and student modeling in quizguide. *User Modeling and User-Adapted Interaction, 25*, 371–424. https://doi.org/10.1007/s11257-015-9164-4

Schubotz, M., Scharpf, P., Teschke, O., Kühnemund, A., Breitinger, C., & Gipp, B. (2020). Automsc: Automatic assignment of mathematics subject classification labels. *Proceedings of the International Conference on Intelligent Computer Mathematics*, 237–250.

Takii, K., Liang, C., & Ogata, H. (2025). Information as interpretation: Measuring learning behavior for knowledge insight. *IEEE Access, 13*, 124197–124210. https://doi.org/10.1109/ACCESS.2025.3583311

Vie, J. J., & Kashima, H. (2019). Knowledge tracing machines: Factorization machines for knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 750–757.

Vovides, Y., Sanchez-Alonso, S., Mitropoulou, V., & Nickmans, G. (2007). The use of e-learning course management systems to support learning strategies and to improve self-regulated learning. *Educational Research Review, 2*(1), 64–74. https://doi.org/10.1016/j.edurev.2007.02.004

Wang, F., King, R. B., & Leung, S. O. (2023). Why do East Asian students do so well in mathematics? A machine learning study. *International Journal of Science and Mathematics Education, 21*(3), 691–711. https://doi.org/10.1007/s10763-022-10262-w

Wang, J., Minematsu, T., Okubo, F., & Shimada, A. (2022). Topic-wise representation of learning activities for new learning pattern analysis. *Proceedings of the 30th International Conference on Computers in Education Conference* (Vol. 1, pp. 268–278).

Xing, W., Liu, Z., Song, Y., Zhu, W., Oh, H., & Li, C. (2025). Development of a generative AI-powered teachable agent for middle school mathematics learning: A design-based research study. *British Journal of Educational Technology, 00*(0), 1–35. https://doi.org/10.1111/bjet.13586

Yamauchi, T., Flanagan, B., Nakamoto, R., Dai, Y., Takami, K., & Ogata, H. (2023). Automated labeling of PDF mathematical exercises with word N-grams VSM classification. *Smart Learning Environments, 10*(1). https://doi.org/10.1186/s40561-023-00271-9

Yamauchi, T., Hoppe, U. H., Dai, Y., Flanagan, B., & Ogata, H. (2024). Representing learning progression of unguided exercise solving: A generalization of wheel-spinning detection. *Proceedings of the 32nd International Conference on Computers in Education* (Vol. 1, pp. 686–695).

Yamauchi, T., Nakamoto, R., Flanagan, B., Dai, Y., Wijerathne, I., & Ogata, H. (2025). Augmentation of learning content with knowledge components: Automatic unit labeling for various forms of Japanese math materials. *IEEE Transactions on Learning Technologies, 18*, 716–731. https://doi.org/10.1109/TLT.2025.3584038