# Educational subject classification with hierarchical information utilizing Large Language Models

Myunghoon Kang°, Jinpyo Hong, Kinam Park*
Department of Computer Science and Engineering, Korea University, (주) Saengcoedu Inc., Human-inspired AI Research
{chaos8527, spkmn}@korea.ac.kr, mindcoding@naver.com

**Abstract**

Humans encode unstructured information hierarchically to understand knowledge effectively and apply it to the real world. Encoding is a core strategy for acquiring knowledge and transforming short-term memory into long-term memory in human cognitive processing. Therefore, building a proper encoding strategy for fostering effective knowledge acquisition is crucial. In this paper, we investigate the possibility of utilizing a large language model to foster human encoding in the educational domain. We first construct an evaluation dataset that consists of middle and high school students' self-study hierarchical summarization and textbook subjects in a pairwise form. In our few-shot experiment, we found that the current large language model shows plausible classification performance when the input is given a hierarchical textual form, implying the possible use case for utilizing the large language model as an educational encoding tutor.

Keywords: Large Language Model, Educational Domain, Hierarchical Information

## 1. Introduction

Human encodes the knowledge into structural form to acquire and store the knowledge into long-term memory [1]. Consequently, developing proper encoding strategy for effective and efficient knowledge acquisition for human has been emerged and studied in education [2, 3, 4]. Moreover, it is important to implement the encoding strategy by personalized learning, since learning is formed through an individual's interaction and experience [5]. Therefore, personalized tutor and effective tutoring for constructing encoding strategy is needed in educational domain.

In the educational domain, various sub-tasks utilize language modeling to realize personalized tutoring with students' products during their learning. Knowledge tracing is a task that predicts the learner's academic achievement, whether a learner solves a given problem based on the learner's problem-solving record [6, 7]. Automated essay scoring ranks or calculates the score, which represents the quality of the student's essay by considering the consistency, clarity, and grammatical errors in the given essay [8]. Despite employing students' record or textual product for personalized learning with the AI, there are no task for building encoding strategy with the students' structural product during learning.

In this work, we present a novel task, educational subject classification with hierarchical information that utilizes large language model (LLM). The task is to measure the processing capability of structural information in educational domain. We construct evaluation dataset that consist of middle and high school student's self-study hierarchical summarization and textbook subject in a pairwise form. Then, we perform experiments with LLM in a few-shot manner for validating the processing capability of structural information in educational domain. We found current large language model shows plausible classification performance implying the possible use case for utilizing large language model as an educational encoding tutor.

## 2. Methods

In this section, we describe the evaluation dataset construction process and the features of the constructed evaluation dataset. Our proposed task, educational subject classification with hierarchical information, consists of a structured keyword set as input and the corresponding type of subject within the textbook as output.

*Corresponding author

<table>
<tr><th>Hierarchical keyword set</th><th>Mapped subject</th></tr>
<tr><td>수의 중국 통일<br>• 수나라<br>  ◦ 문제<br>    ▪ 업적<br>      ▪ 남북조 통일<br>      ▪ 과거제 실시<br>      ▪ 북위 이래 제도 정비<br>        ▪ 예<br>          ▪ 토지<br>          ▪ 조세<br>          ▪ 군사<br>    ▪ 효과<br>      ▪ 중앙 집권 체재 강화</td><td>세계사 (world history)</td></tr>
<tr><td>▪ 우주의 구성원소<br>  ▪ 주요 원소<br>    ▪ 종류<br>      ▪ 수소<br>      ▪ 헬륨<br>    ▪ 특징<br>      ▪ 전체의 98%를 차지<br>▪ 지구의 구성원소<br>  ▪ 주요원소<br>    ▪ 철</td><td>과학 (science)</td></tr>
</table>

Figure 1. Detailed example of the evaluation dataset

| **System prompt** |
|---|
| 당신에게 입력으로 문자열(str)로된 markdown 파일이 주어질 것입니다. 주어진 markdown 파일이 속한 교과 과목을 선택지에서 골라주세요.<br>내용이 여러 과목에 걸쳐 있어도 가장 적절한 한 개의 과목만을 선택해주세요. 예측 이유에 대한 설명은 하지마세요.<br><br>선택지: a) 사회, b) 역사, c) 과학, d) 비문학<br>markdown 입력: {input}<br>정답: {output}<br>markdown 입력: {input}<br>정답: {output} |

Table 1. 2-shot prompt used for the experiment

To construct the dataset, we gathered four students' self-study textbook encoding output, varying from middle school to high school. As shown in Figure 1, students extract the core keywords and list the keywords in structured and hierarchical form. Since the original encoding output is merged into a single markdown file, we divide the file into unit-level sections to evaluate the language model's structural processing at a fine-grained level.

The statistics of the proposed evaluation dataset is shown in Table 2. The dataset consists of four subjects, covering a variety of educational domain: science, non-literature, social studies, and world history.

| **Subject** | **Number of dataset** |
|---|---|
| Non-literature | 21 |
| Science | 26 |
| Social studies | 34 |
| World history | 28 |
| Total # | 109 |

Table 2. Statistics of the proposed evaluation dataset

## 3. Experiment

We conduct an experiment to verify the LLM's structural comprehension capability in the educational domain. For the experiment, we use llama-3-Korean-Bllossom-8B*, Korean instruction-tuned model based on Llama3 [9]. Given the student's self-study hierarchical information, we prompt the model to predict the subject type. The prompt used for the experiment is shown in Table 1. We use social studies and science for the 2-shot example. For the evaluation metric, we use precision, recall, and f1-score to assess the structural

*https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B

| Model | Subject | Precision | Recall | F1 |
|---|---|---|---|---|
| llama-3-Korean-Bllossom-8B | Non-literature | 64.00 | 76.00 | 70.00 |
| | Social studies | **92.00** | 68.00 | 78.00 |
| | World history | 89.00 | 86.00 | **87.00** |
| | Science | 78.00 | **96.00** | 86.00 |
| | Overall | 81.00 | 81.00 | 80.00 |

Table 3. Education subject classification result with hierarchical information

comprehension capability in a classification manner.

The main result is presented by the subject in Table 3. The llama-3-Korean-Bllossom-8B showed a plausible result, achieving 80% in the F1-score. Regarding precision and recall, both social studies and science achieved the best results in the subject. This is due to the 2-shot example's domain, which gave a marginal hint for comprehensing the structural form of the student's hierarchical information. Interestingly, the llama-3-Korean-Bllossom-8B model demonstrated the best performance in F1-score for the world history subject, despite this subject not receiving any significant cues from the 2-shot examples. However, the model showed lowest performance on the non-literature subject, showing 67% in precision. In conclusion, our result shows the possibility of utilizing LLM for comprehensing structural information in educational domain, but future work is needed to enhance the comprehensibility of the LLM in processing hierarchical information.

## 4. Conclusion

In this paper, we propose a novel task of educational subject classification with hierarchical keyword information from students' self-study output to support learners' building encoding strategy. To this end, we constructed an evaluation dataset and evaluated the LLM's hierarchical information comprehensibility in the educational domain. In future research, we plan to develop an effective training method for enhancing LLM's hierarchical information comprehensibility in the educational domain.


## Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Culture, Sports and Tourism R&D Program through the korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name : Development of generative AI-based publishing content analysis and sharing platform technology to respond to changes in the publishing environment. Project Number : : RS-2024-00442061, Contribution Rate : oo%). Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.


## Reference


[1] R. C. Atkinson, "Human memory: A proposed system and its control processes," *The psychology of learning and motivation*, Vol. 2, 1968.

[2] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, Vol. 63, No. 2, p. 81, 1956.

[3] S. B. Rimm, *When gifted students underachieve: What you can do about it.* PRUFROCK PRESS INC., 2006.

[4] S. Schellenberg, M. Negishi, and P. Eggen, "The effects of metacognition and concrete encoding strategies on depth of understanding in educational psychology." *Teaching Educational Psychology*, Vol. 7, No. 2, pp. 17–24, 2011.

[5] A. Shemshack and J. M. Spector, "A systematic literature review of personalized learning terms," *Smart Learning Environments*, Vol. 7, No. 1, p. 33, 2020.

[6] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, Vol. 4, pp. 253–278, 1994.

[7] N. Liu, Z. Wang, R. Baraniuk, and A. Lan, "Open-ended knowledge tracing for computer science education," *Proceedings of the 2022 Conference on Empirical*

*Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., pp. 3849–3862, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.emnlp-main.254

[8] Z. Jiang, T. Gao, Y. Yin, M. Liu, H. Yu, Z. Cheng, and Q. Gu, "Improving domain generalization for prompt-aware essay scoring via disentangled representation learning," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12 456–12 470, 2023.

[9] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.