Leveraging LLMs for Adaptive Testing and Learning in Taiwan Adaptive Learning Platform (TALP)

Bor-Chen Kuo [1], Frederic T. Y. Chang [1] and Zong-En Bai [1]

[1] National Taichung University of Education, 140 Minsheng Rd., West Dist., Taichung City, 403514, Taiwan

Abstract

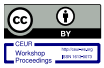
Artificial Intelligence (AI) and Large Language Models (LLMs) have gained prominence in the educational context, revolutionizing various aspects of teaching and learning. This study focuses on the feasibility of integrating LLMs into the Taiwan Adaptive Learning Platform (TALP) to improve its current adaptive mechanism and enhance the learning experience of students. Through an in-depth exploration, the study identifies several potential benefits of incorporating LLMs into TALP. Firstly, by harnessing the power of LLMs and combining them with the existing knowledge structure in TALP, qualitative responses from open-ended questions can be analyzed more effectively. This enables a more precise assessment of students' understanding and significantly reduces the number of unnecessary testing items, saving valuable time and resources. Additionally, the integration of a chatbot in TALP's diagnostic report provides an innovative approach to scaffolding during remediation. The chatbot can engage in Socratic interactions with students, guiding them through the learning process and addressing misconceptions in real-time. This personalized support fosters a deeper understanding of the material and facilitates more effective remediation. Furthermore, the study highlights the potential of LLMs in detecting and addressing individual learning weaknesses. By leveraging the deep interaction capabilities of LLMs, TALP can analyze student responses and identify cross-grade misconception more efficiently. Finally, the implementation of LLMs in TALP also presents challenges which is discussed. In conclusion, integrating LLMs into TALP holds great potential to enhance its adaptive mechanism, provide personalized learning experiences, and address individual learning weaknesses.

Keywords

LLMs, Adaptive learning, Chatbot, Learning platform, GPT

1. Introduction

Artificial Intelligence (AI) has become increasingly involved in the educational context with the introduction of Large Language Models (LLMs), particularly since the advent of ChatGPT. Several studies have sought to apply LLMs in education for various purposes, including tutoring and homework assistance, language learning, writing assistance, personalized learning, and interactive learning [1, 2, 3]. At present, it is difficult to predict whether LLMs like ChatGPT, or their subsequent versions, will fully replace teachers. However, we are more interested in exploring how the application of LLMs can enhance the effectiveness of current educational tools. The Taiwan Adaptive Learning Platform (TALP) is the official learning platform of the Ministry of Education (MOE) in Taiwan, serving 2.8 million registered users from grades 1 to 12. A unique feature of TALP is its use of AI to provide individual learning paths for personalized learning. According to a large-scale survey by MOE Taiwan, this platform is highly effective in enhancing students' academic achievement and promoting self-regulated learning [4]. In this study, we will explore the feasibility of introducing LLMs to TALP and investigate whether its implementation can improve TALP's current adaptive mechanism.

Proceedings Name, Month XX–XX, YYYY, City, Country
EMAIL: email1@mail.com (A. 1); email2@mail.com (A. 2); email3@mail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)


2. The application of LLMs in TALP

In the following four sections, our study will provide an introduction to the current adaptive mechanism in TALP. We will then propose the implementation of LLMs technology to enhance the effectiveness of adaptive testing and learning in TALP. Finally, we will discuss the potential obstacles that may arise during the implementation of LLMs in TALP.

2.1. The Current adaptive mechanism in TALP

The conceptual framework of the adaptive mechanism in TALP is to: apply adaptive tests for diagnosing learning weakness and; offer an individual learning path based on the result of diagnosis for fixing learning mistakes. The current adaptive testing of TALP is to apply rule-based AI technology judged by the response of multiple choice items from test takers. For example, as shown in Figure 1(a), the testing system will select a question related to the highest-level concept (A) from the question database for the test taker to answer. If the test taker answers question A incorrectly, according to rule-base, the testing system will then choose questions related to the lower-level concepts (B and C) for testing. If the test taker answers question B correctly, it is then predicted that they would answer the sub-concepts of B (D and E) correctly as well, so there is no need for them to answer these questions. However, if the test taker answers question C incorrectly, the system will then select questions related to the sub-concepts of C (F, G, and H) for the test taker to continue answering.

As highlighted by Wu, Kuo, & Yang (2012), the application of knowledge structure with ordering theory AI algorithms to diagnostic tests carries several advantages, such as: 1. tracing learning paths across students; 2. visualizing learning paths; 3. saving unnecessary test items during diagnosis. Wu, Kuo & Wang, (2017) have demonstrated that high levels of effectiveness and efficiency of knowledge structure which may increase the accuracy of identifying learning weaknesses by up to 90%, meanwhile saving 80% of unnecessary items during testing.

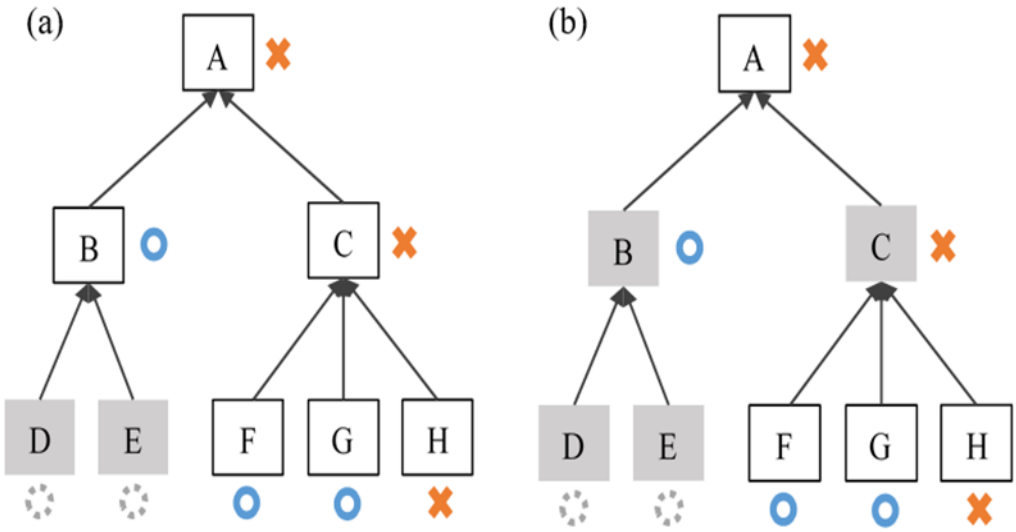


Figure 1: Comparison of adaptive testing AI system (a) rule-base (b) LLMs

Knowledge structure in TALP is presented as shown in Figure 2(a), it is similar to that of a sky map which is made of knowledge nodes. When students complete an adaptive diagnostic test in TALP, the result will reflect on the color of notes and sub-notes in the knowledge structure. Notes in green indicate that students have mastered the skills; ones in orange reveal that students fail to master the skills. The individual learning path is plotted by connecting orange notes in the knowledge structure (as shown in Figure 2(a)). In TALP, each subskill contains an instructional video, in-video quizzes, exercise and dynamic assessments for fixing mistakes. Once students competently complete watching the videos and

pass the tests, the color of the nodes will turn green. The Learning path also can be transferred into the diagnosis report shown in Figure 2(b), it indicates not only the progress in the learning path, but also the percentage of completion in instructional video, quizzes, exercise and dynamic assessment.

Figure 2: (a) personalized learning paths in TALP; (b) diagnosis report

2.2. Elevating the efficiency in diagnosing learning weakness by LLMs

In the past, multiple-choice items were the preferred format for computer-aided testing due to their straightforward nature (right or wrong). Evaluating open-ended questions, which provide more qualitative and richly informative responses, posed a significant challenge due to the limitations of computer technology [5]. The advantage of LLMs is able offer the service of automated response analysis which can examine and evaluate the response of open-ended questions from test takers. As the open-ended questions can reveal the response including arithmetic process that can offer rich information for LLMs to judge misconceptions directly. For example, if concept A in Figure 1 involves the four arithmetic operations, concept B refers to arithmetic operations involving addition and subtraction and concept C indicates arithmetic operations in multiplication and division. In the present TALP system, if the test taker answers a question incorrectly in concept A, the system will provide two items related to concept B and C, respectively. Due to the nature of an open-ended question, the answer includes the calculation process, which can effectively demonstrate the level of mastery in arithmetic. However, it is important to consider the scenario where a student excels in addition and subtraction but struggles with multiplication and division, as illustrated in Figure 1(b). TALP intends to incorporate LLMs (Language Model Models) technology into adaptive testing, enabling it to assess students' responses to open-ended questions similar to how a teacher would evaluate them. In the scenario depicted in Figure 1, the utilization of LLMs in the TALP adaptive testing system has the potential to save an additional two items.

2.3. Better Scaffolding and diagnosis in remediation by LLMs

Though the current learning resource in TALP is bountiful with instructional videos and assessment module to remediate the learning weakness displayed in the diagnostic report. Vygotsky's sociocultural theory identified that learning is an essentially social process, it is vital to learn through the guidance of teachers or the collaboration with peers [6,7]. Many researchers have made efforts to apply technology for simulating tutors, such as intelligent tutor systems (ITS)[8]. However, the level of engagement and feedback provided by these systems has still remained unsatisfactory. The current diagnostic report of TALP could list the desired learning material for individual, but it can't provide instant feedback anytime during remediation. By utilizing LLMs technology, implementing chatbot in the diagnostic report of TALP can interact with students while remediating their learning weakness. As shown in Figure 3, the TALP system prompts GPT with relevant remedial information. This includes the learning content in the knowledge structure along with its corresponding notes, related testing items, and previous testing responses. The above prompt aims to help GPT understand students more and provide feedback with better quality. In the planned diagnostic report of TALP, chatbot for remediation is optional. Once students use chatbot for learning, TALP will open a chat box where students can do their quizzes or assessment with chatbot. In our settings, chatbot will intact with students by Socratic methods, instead of providing direct answers, the chatbot uses probing questions to guide students in discovering knowledge, examining their performance, and engaging in logical reasoning. Once students have completed the remedial tasks scaffolded by the chatbot, the TALP system collects the dialogue information. This information is then utilized by GPT to generate customized assessment items, specifically targeting the individual student's learning weaknesses. The purpose is to assess and evaluate whether students have achieved a thorough mastery of the required competence. As depicted in Figure 3, when students answer correctly, the TALP system will guide them in remediating higher-level misconceptions. When students are unable to provide a correct answer, the chatbot guides them either back to the original remedial process or towards a more in-depth remediation at a lower level. Though the current diagnostic system in TALP is cross-grade with good accuracy, through the deeper interaction with students, LLMs can help rectify diagnostic errors that may occur in the original one.

Within the region delineated by the red rectangle in Figure 3, we have designated a space for students to engage with the chatbot. Each competence listed in the diagnostic report has its own activation button for the chatbot, as our tests with GPT-3.5 showed that it could not completely resolve all issues in the testing bank. In our preliminary trial, GPT3.5 managed to correctly answer just around 70% of the fifth-grade level questions from the TALP testing bank. A significant portion of errors originated from misinterpretations of Chinese mathematical terminology and symbols. Indeed, many of these errors can be mitigated largely through the use of well-constructed prompts beforehand or by opting for GPT-4. At present, the TALP research team favors the use of prompts over employing GPT-4. This preference is driven by the limited availability of the GPT-4 API, coupled with its higher cost compared to GPT3.5, which represents substantial savings for TALP. Looking ahead to the end of July, a selection of chat interfaces will be equipped with customized prompts in the diagnostic report of TALP aimed at improving GPT-3.5's understanding. This is expected to augment the feedback provided by the chatbot, thereby bolstering the problem-solving capabilities of students during interactions.

2.4 The challenges in implementing LLMs in TALP

As the previous discussion shown, LLMs is so potential to improve the current adaptive mechanism in TALP. Combining with the knowledge structure, TALP equipped with LLMs technology may save more unnecessary items than before. Integrating a chatbot into the diagnostic report can create an improved scaffold for remediation, facilitated by Socratic interactions. Additionally, the chatbot's deep interaction capabilities can enable further diagnosis of the student's understanding. To accomplish the aforementioned objectives, the achievement relies on the accuracy and precision of GPT in interpreting and providing answers to the learning content and testing items, especially in mathematics and science. Evidently, the API of GPT-3.5 is accessible for constructing chatbots within the platform. However, its precision and accuracy in problem-solving and interpretation of mathematical symbols are areas that still require improvement [9]. Even if GPT-4 were currently available, while it may offer improved

accuracy and precision, the problem-solving capabilities of GPT-4 would still need to be demonstrated and validated. The cost associated with GPT-4 poses a significant challenge that needs to be addressed, particularly for operating a learning platform like TALP, which is fully funded by the Ministry of Education (MOE) and renowned for providing free usage to grade 1-12 students.

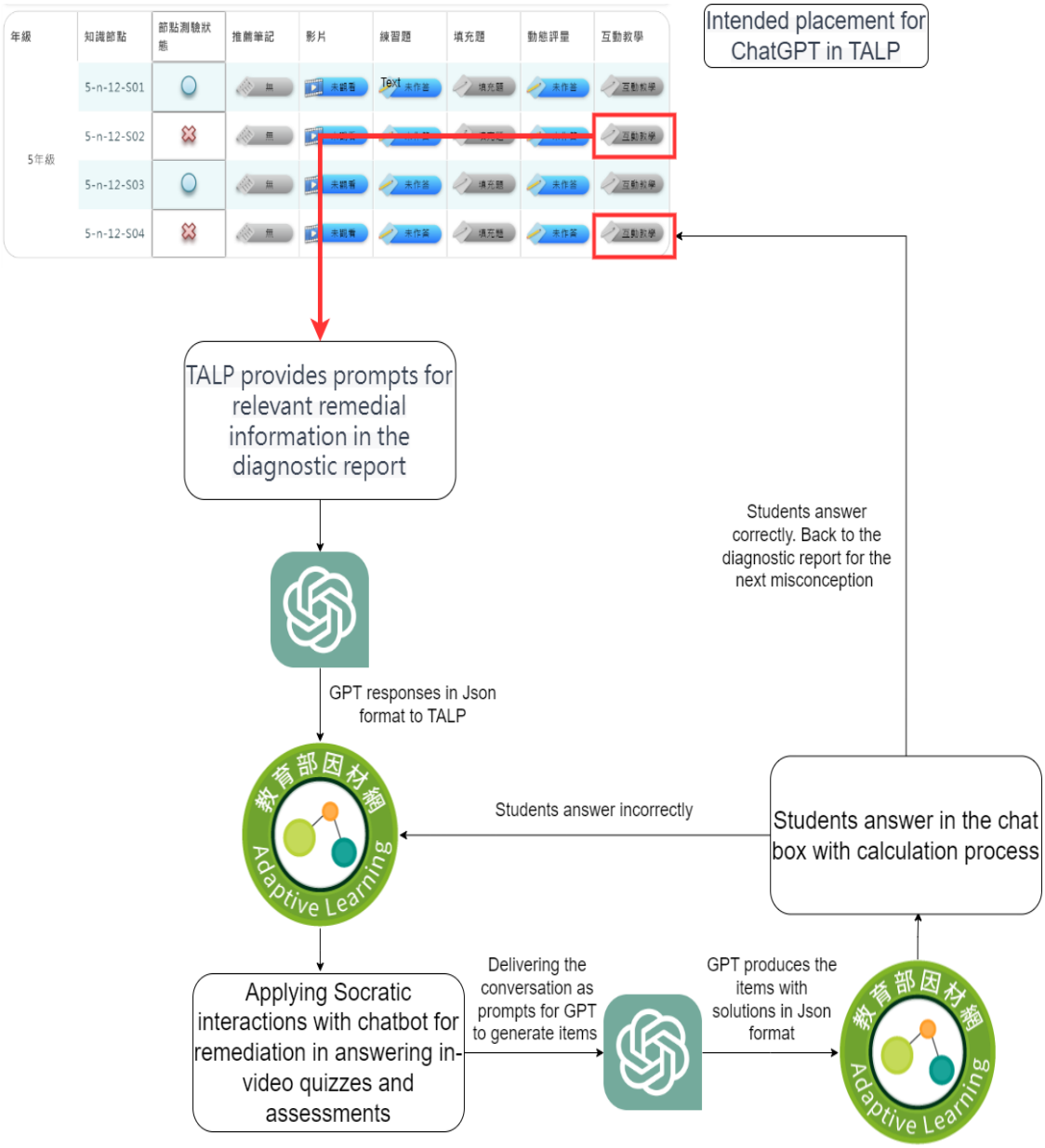


Figure 3: Implementing chatbot in the TALP's diagnostic report

3. Acknowledgements

We gratefully acknowledge the financial support and backing of both the Ministry of Education and the National Science and Technology Council of Taiwan. Their generosity and commitment have been instrumental in the advancement of this study. We extend our heartfelt appreciation for their steadfast belief in our research endeavors.

4. References

[1] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, p. 102274, 2023/04/01/ 2023, doi: https://doi.org/10.1016/j.lindif.2023.102274.

[2] M. Fraiwan and N. Khasawneh, "A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions," arXiv preprint arXiv:2305.00237, 2023.

[3] W. Dai et al., "Can large language models provide feedback to students? A case study on ChatGPT," 2023. doi:10.35542/osf.io/hcgzj.

[4] Huang, H.-Y. (2022, December 14). Analyzing the Effectiveness of the Taiwan Adaptive Learning Platform on Learning Outcomes using Educational Data and Data Mining Techniques. Paper presented at the 2022 Self-Regulated Learning Festival and Learning Analytics Seminar, Kaohsiung, Taiwan.

[5] M. Stevenson, "Shermis, MD, & Burstein, J.(Eds)(2013). Handbook of Automated Essay Evaluation: Current applications and new directions," Journal of Writing Research, vol. 5, no. 2, pp. 239-243, 2013.

[6] S. McLeod, "Vygotsky's Zone of Proximal Development and Scaffolding", 2023. URL:https://www.simplypsychology.org/zone-of-proximal-development.html?ref=brainscape-academy.

[7] L. S. Vygotsky and M. Cole, Mind in society: Development of higher psychological processes. Harvard university press, 1978.

[8] K.-C. Pai, B.-C. Kuo, C.-H. Liao, and Y.-M. Liu, "An application of Chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning," Educational Psychology, vol. 41, no. 2, pp. 137-152, 2021.

[9] X. Liu, T. Pang, and C. Fan, "Federated Prompting and Chain-of-Thought Reasoning for Improving LLMs Answering," arXiv preprint arXiv:2304.13911, 2023.