

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

Bor-Chen Kuo<sup>1</sup>, Frederic T. Y. Chang<sup>1</sup> and Zong-En Bai<sup>1</sup>

<sup>1</sup> 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 into TALP's diagnostic report provides an innovative approach for 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 misconceptions more efficiently. This study also provides examples of how GPT-3.5 can be applied for the above purposes. Finally, the implementation of LLMs in TALP also presents challenges, which are 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


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


---

Empowering Education with LLMs – the Next-Gen Interface and Content Generation, July 7, 2023, Tokyo, Japan

 kbc@mail.ntcu.edu.tw (B. Kuo)

 0000-0003-1741-2450 (B. Kuo)

 © 2023 Copyright for this paper by its authors.  
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

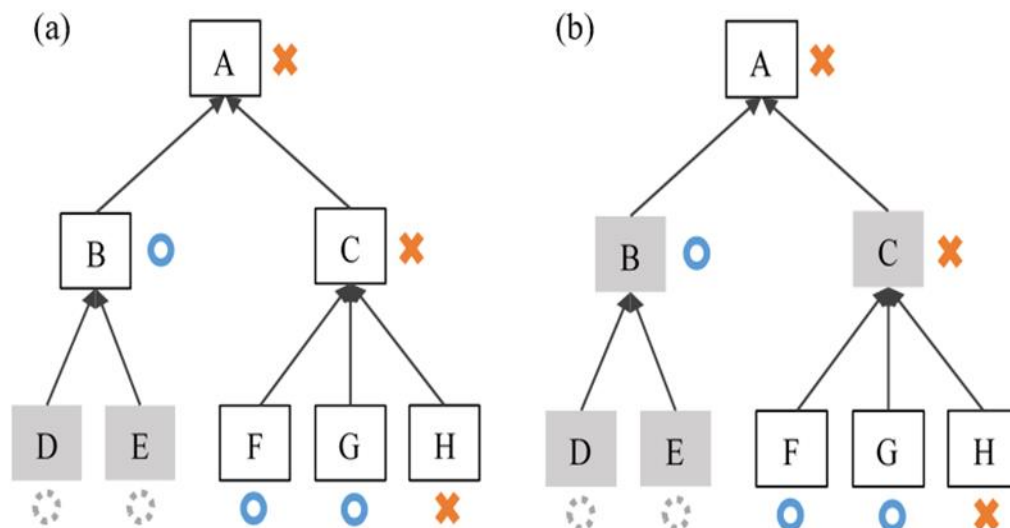
## 2. The application of LLMs in TALP

In the following four sections, our study will delve into the existing adaptive mechanism of TALP. Subsequently, we will propose the implementation of LLM technology to augment the efficacy of adaptive testing and learning within TALP. We will also present some examples of how LLMs could be applied in TALP to facilitate learning. Finally, we will discuss potential obstacles that might arise during the integration of LLMs into TALP.

### 2.1. The Current adaptive mechanism in TALP

The conceptual framework of the adaptive mechanism in TALP involves two main steps: applying adaptive tests to diagnose learning weaknesses, and offering an individual learning path based on the diagnosis to remedy learning mistakes. The current adaptive testing in TALP applies rule-based AI technology, which is guided by the responses of test-takers in multiple-choice items. 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, the testing system, based on the rule-base, will then select questions related to the lower-level concepts (B and C) for further 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 subsequently present questions related to the sub-concepts of C (F, G, and H) for the test taker to answer.

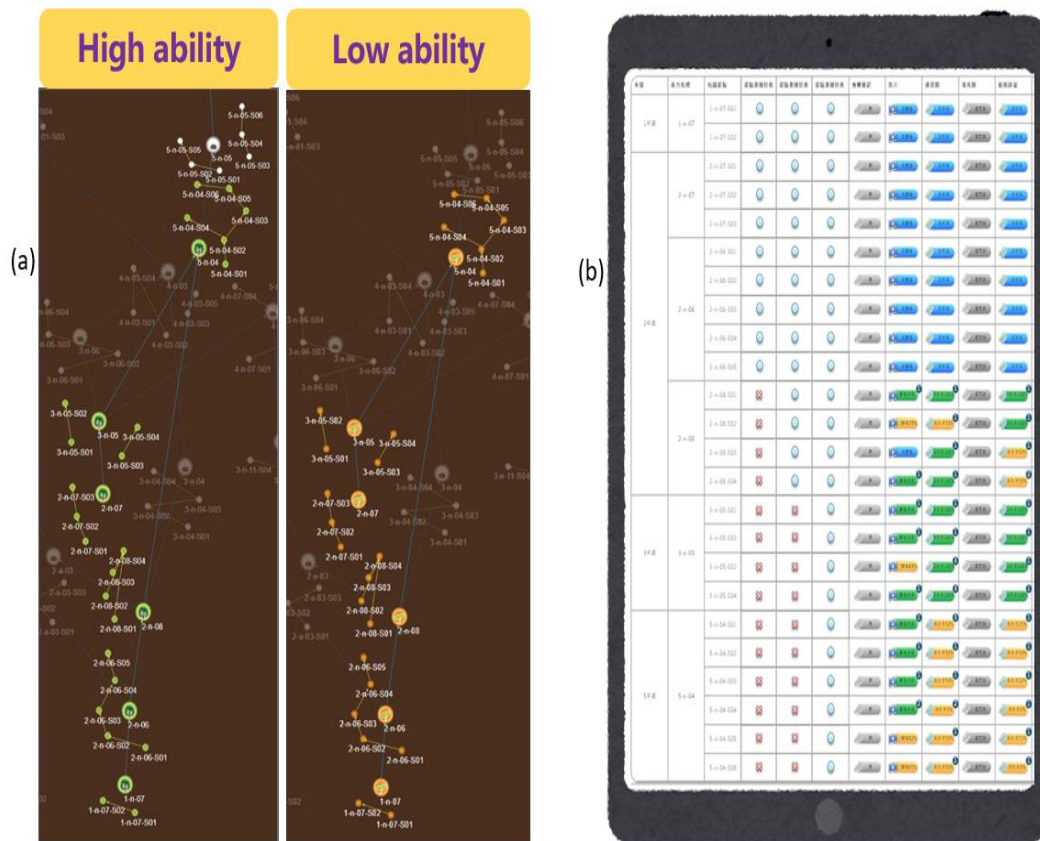
As highlighted by Wu, Kuo, & Yang (2012), employing AI algorithms that incorporate knowledge structure with ordering theory in diagnostic tests offers several advantages, including: 1. tracing learning paths across students, 2. visualizing learning paths, and 3. eliminating unnecessary test items during diagnosis. Wu, Kuo & Wang (2017) have demonstrated that the high effectiveness and efficiency of knowledge structure can increase the accuracy of identifying learning weaknesses by up to 90%, while simultaneously reducing up to 80% of unnecessary items during testing.



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

The knowledge structure in TALP, as shown in Figure 2(a), resembles a sky map composed of knowledge nodes. When students complete an adaptive diagnostic test in TALP, the results are reflected in the color of the nodes and sub-nodes in the knowledge structure. Nodes colored green indicate that students have mastered the skills, while those in orange reveal areas the students have yet to master. The individual learning path is plotted by connecting the orange nodes in the

knowledge structure, as shown in Figure 2(a). Each subskill within TALP includes an instructional video, in-video quizzes, exercises, and dynamic assessments aimed at correcting mistakes. Once students competently complete watching the videos and pass the tests, the color of the nodes turns green. The learning path can also be converted into a diagnostic report as illustrated in Figure 2(b). This report not only indicates progress along the learning path but also displays the percentage of completion for the instructional video, quizzes, exercises, 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 their ability to offer a service for automated response analysis, which can examine and evaluate the responses to open-ended questions from test-takers. Open-ended questions can reveal arithmetic processes, providing rich information for LLMs to directly identify misconceptions. 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 current TALP system, if the test taker answers a question related to concept A incorrectly, the system will provide two items related to concepts B and C, respectively. Due to the nature of open-ended questions, the answer includes the calculation process, which can effectively demonstrate the level of mastery in arithmetic. However, it's important to consider scenarios where a student excels in addition and subtraction but struggles with multiplication and division, as illustrated in Figure 1(b). TALP plans to incorporate LLM technology into adaptive testing, enabling it to assess students' responses to open-ended questions similarly 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.

### **3. Better Scaffolding and diagnosis in remediation by LLMs**

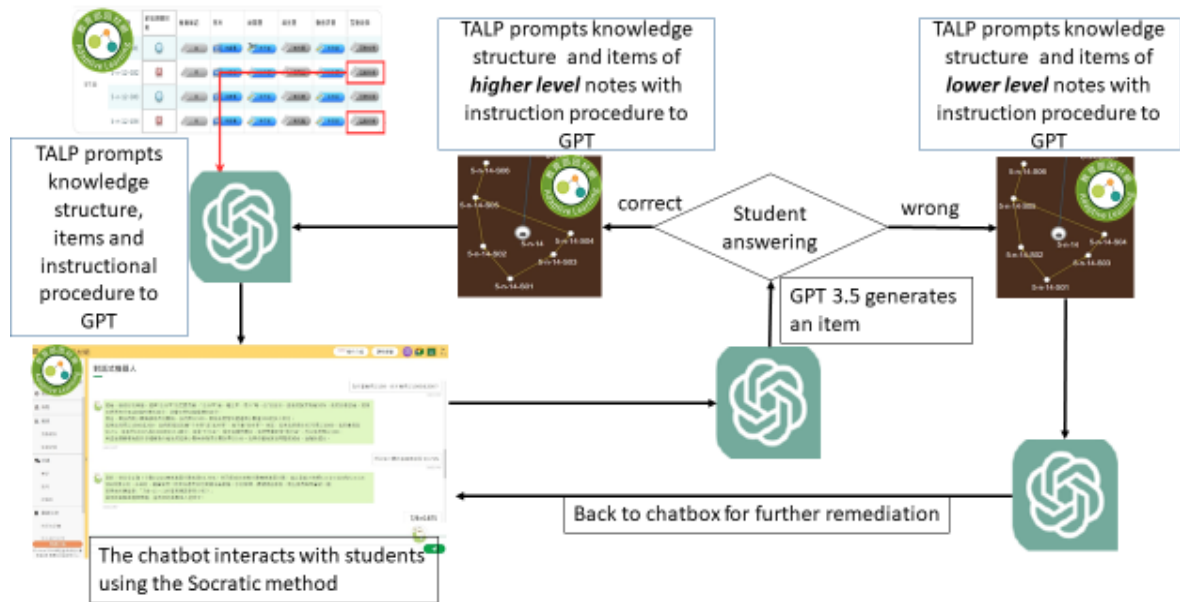
Although the current learning resources in TALP are abundant, with instructional videos and the assessment module to remediate learning weaknesses in the diagnostic report, the importance of social processes in learning should be addressed. As identified in Vygotsky's sociocultural theory, learning is essentially a social process; guidance from teachers or collaboration with peers is vital [6,7]. Many researchers have endeavored to simulate tutors using technology, such as Intelligent Tutoring Systems (ITS)[8]. However, the level of engagement and feedback provided by these systems has remained unsatisfactory. The current diagnostic report of TALP can list the desired learning material for an individual, but it cannot provide instant feedback anytime during remediation.

Significant improvements were not realized until the advent of Large Language Models (LLMs). Some researchers have utilized BERT to solve mathematical problems and attempted to provide students with feedback by assessing their responses. In the realm of automatic item generation, researchers historically relied on item templates [10], but with the introduction of LLMs, some have begun to generate items based on students' responses to test questions [11]. While early LLMs may have made some progress in these areas, it is noteworthy that no LLMs were able to integrate the above tasks, especially in the context of Traditional Chinese. However, with the advent of GPT-3.5, it has become feasible to implement a chatbot in TALP's diagnostic report that can interact with students while addressing their learning weaknesses.

Figure 3 illustrates how the TALP system utilizes GPT, providing it with pertinent remedial information. These prompts are critical to successfully diagnosing mathematical problems, interacting with students, offering feedback and instructions, as well as generating test items. In our design, we employ the framework presented in [10] for our prompts, which encompasses a cognitive model, an item model, and an instructional procedure. The cognitive model represents the knowledge structure along with its associated notes. Meanwhile, the item model refers to the test items, and the instructional procedure outlines how GPT will engage with students. 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 the settings of instructional procedure, chatbot will interact 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 towards a more in-depth remediation at a lower level. While the existing diagnostic system in TALP offers cross-grade precision with commendable accuracy, we believe integrating it with LLMs could further improve the results. By enabling deeper interaction with students, the collaboration of LLMs with the current rule-based TALP AI system could offer more nuanced diagnostics of learning weaknesses.

Of the numerous large language models available on the market, GPT-3.5 emerges as our top choice, especially due to its efficient and fluent handling of Traditional Chinese content. This selection was also necessitated by the current unavailability of GPT-4 for TALP. To assess its capabilities, we conducted an initial test using 569 5th-grade test items derived from TALP. The mathematical problem-solving accuracy rate was 79% initially, which, while promising, indicated

potential for further enhancement. By integrating both the cognitive and item model into the prompts, the accuracy witnessed a significant rise to 96%. Intriguingly, when we employed GPT-4 prompting with both the cognitive model and item model, the accuracy impressively peaked at 100%. The above results suggest that while GPT-4 stands out as the superior engine, GPT-3.5 can also deliver comparable outcomes when provided with carefully structured prompts, thereby effectively meeting our requirements.



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

#### 4. Examples by applying GPT-3.5

In the following examples, we will illustrate the feasibility of employing GPT-3.5 to enhance TALP's diagnostic reports in mathematics by: (1) pinpointing the students' learning weaknesses and saving testing items; (2) scaffolding their learning through interaction with a chatbot; and (3) generating assessment items. The domain knowledge pertains to fifth-grade level understanding of ratios and their practical applications in everyday life, encompassing concepts such as 'percentage' and 'discount'.

To achieve the aforementioned goals, the input of appropriate prompts is crucial for the success of our task. Two prompts are required: one for the assessment item and domain knowledge, and another for the instructional procedure. Given that our task involves automated item generation and automated rating, it is imperative to have well-defined cognitive and item models [10], as outlined in Table 1, to clarify the testing domain knowledge. Table 1 also shows how indicators of knowledge structure are utilized for the cognitive model.

In the section related to the item model, we input data as multiple-choice questions, comprising stems, options, and answers. This comprehensive information significantly aids GPT-3.5 in understanding the context of tests. Our initial trials have shown that structuring prompts in multiple-choice format can substantially improve the accuracy of the feedback provided. Additionally, the distractors included in multiple-choice items serve to effectively illustrate common misconceptions to GPT-3.5.

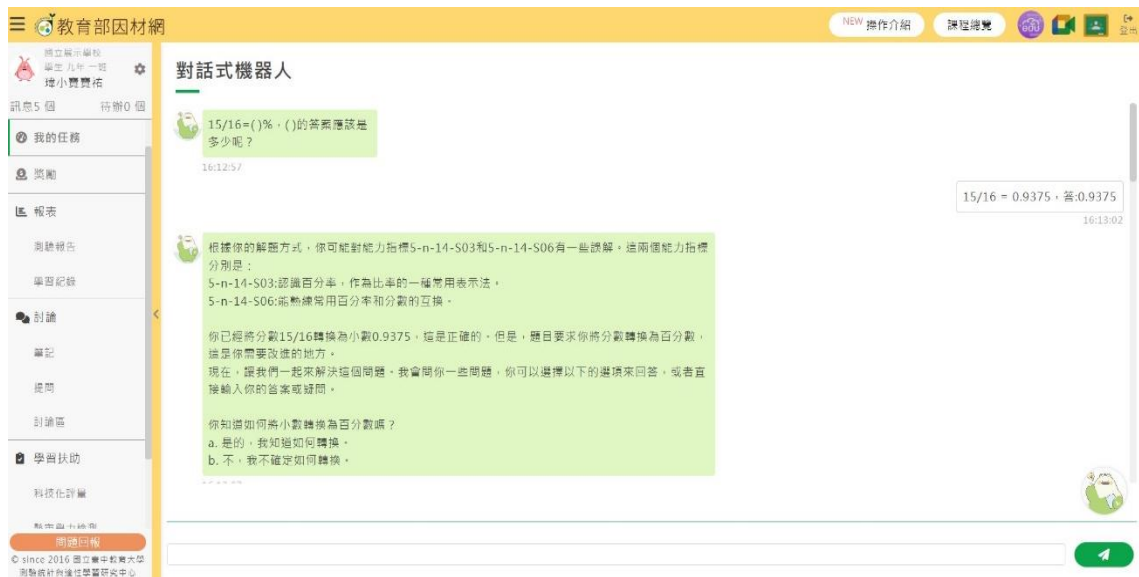
Lastly, establishing an instructional procedure is essential for the chatbot to effectively guide students through the Zone of Proximal Development (ZPD). A Socratic interaction approach is applied to scaffold and guide students towards understanding. The prompt for the instructional procedure can be found in Table 1.

**Table 1**  
**Cognitive model and item model for prompts**

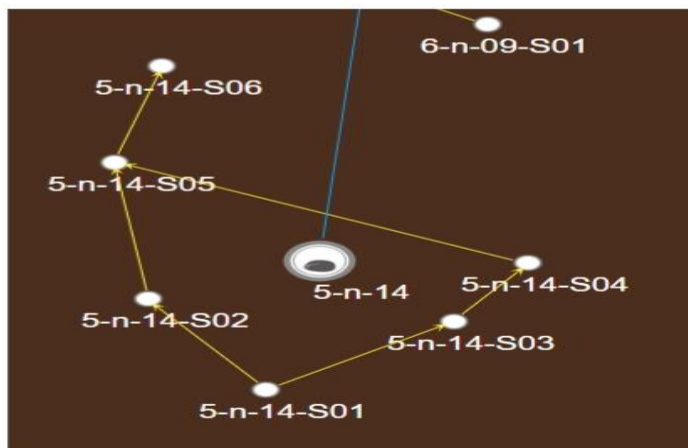
Cognitive model
Knowledge Structure Note 5-n-14 5-n-14-S01: Understand the concept of ratio as "the amount of a part compared to the total." 5-n-14-S02: Able to solve problems related to ratios in daily life. 5-n-14-S03: Understand percentages as a commonly used representation of ratios. 5-n-14-S04: Able to solve problems related to percentages in daily life. 5-n-14-S05: Able to solve applied problems related to percentages in daily life (including discounts and increases). 5-n-14-S06: Proficient in converting between commonly used percentages and fractions. The hierarchical knowledge structure in 5-n-14 is as follows: At the top is S06, which is preceded by S05. S05 then precedes both S04 and S02. Continuing down the hierarchy, S04 is above S03, while S03 and S02 are both positioned over S01.
Item model
Stem: "15/16=( )% , ( )的答案應該是多少呢?" Option: (1) 93.75 (2) 9375 (3) 0.9375 (4) 9.375 Answer: (1)
Instructional procedure
Prompts to GPT3.5: Analyze students' mistakes using indicators of knowledge structure, and pinpoint which area they are struggling with. Subsequently, employ scaffolding. Rather than directly providing the correct answer, use the Socratic method to guide students in thinking and explaining. Based on the students' responses, offer explanations and guidance. Depending on the student's response, generate a testing item similar to the one in the item model to assess the level of learning. If the answer is correct, it is assumed that the student has grasped the material; if the answer is incorrect, continue providing guidance until the correct answer is given.

In Figure 4(a), the chatbot displays a question labeled as 5-n-14-S06 (shown in Figure 4(b)) in the knowledge structure for the student to solve. Based on the students' responses, the chatbot identifies that their learning weakness, attributed to failing 5-n-14-S06, is rooted in 5-n-14-S03. In the knowledge structure, 5-n-14-S03 is the competence of understanding percentages as a commonly used representation of ratios. In the other word, it refers to convert a decimal into a percentage. In contrast, the previous TALP AI rule-based diagnostic system would require testing the students on 5-n-14-S05, 5-n-14-S04, and 5-n-14-S02 to pinpoint the actual learning deficiency located in 5-n-14-S03. Employing GPT-3.5 as an automated rater streamlines the diagnostic process by reducing the number of test items needed.





**Figure 4 (a):** Implementing chatbot as automated rater in Diagnostic report



**Figure 4(b):** The knowledge structure of 5-n-14 (understanding of ratios and their practical applications in everyday life)

In the chatbox (as depicted in Figure 5), the chatbot interacts with students by providing instructions for remediation. In this instance, the student was identified as having difficulty converting a decimal to a percentage. The chatbot sought to teach the student how to convert a decimal into a percentage. It demonstrated this by showing that multiplying a decimal by 100 and appending the percentage symbol to the result accomplishes the conversion. As shown in Figure 5, though the chatbot demonstrated the method of converting a decimal into a percentage, the student still had doubts regarding this demonstration. To address these doubts, the chatbot, within the chatbox, used the Socratic method to guide the student towards understanding the concept. This was achieved by providing additional explanations and posing simple questions for the student to answer. After completing the instruction, the chatbot generates a testing item similar to the original one in Figure 4(a) to assess whether the student has mastered the concept (as seen in Figure 6). This process aims to test whether students can complete tasks independently without assistance from the chatbot.

教育部因材網 NEW 操作介紹 課程總覽

國立屏東師範大學 學生九年一貫 潘小賢寶祐

訊息5個 待辦0個

我的任務

獎勵

報表

測驗報告

學習紀錄

討論

筆記

提問

討論區

學習扶助

科技化評量

問題回報

© since 2016 國立屏東教育大學 測驗統計與進修學習研究中心

### 對話式機器人

我不確定如何將小數轉換為百分數 16:13:12

沒問題，我們可以一起學習這個概念。  
將小數轉換為百分數的基本步驟是將小數乘以100，然後在數字後面加上百分號(%)。這是因為“百分”的意思是“每一百”，所以我們將小數乘以100來找出它是“每一百”的多少。  
現在，讓我們試著將你的小數0.9375轉換為百分數，我們應該進行哪一步操作呢？  
A.  $0.9375 \times 10$   
B.  $0.9375 \times 100$   
C.  $0.9375 / 100$   
D.  $0.9375 \div 100$

請選擇你認為正確的選項，或者如果你有其他疑問，也可以直接提出。  
16:13:17

B 16:13:22

非常好！你選擇的選項是正確的。我們將小數0.9375乘以100來將其轉換為百分數。  
所以，我們的計算會是： $0.9375 \times 100 = 93.75$ 。

教育部因材網 NEW 操作介紹 課程總覽

國立屏東師範大學 學生九年一貫 潘小賢寶祐

訊息5個 待辦0個

我的任務

獎勵

報表

測驗報告

學習紀錄

討論

筆記

提問

討論區

學習扶助

科技化評量

問題回報

© since 2016 國立屏東教育大學 測驗統計與進修學習研究中心

### 對話式機器人

為什麼是乘以100，而不是乘以1000或200？ 16:13:32

這是一個很好的問題。這與“百分率”的定義有關。“百分率”是一種比率，表示“每一百”的部分。當我們說某物是50%，我們的意思是，如果你將某物分成100個相等的部分，那麼你有50個這樣的部分。  
因此，當我們將小數轉換為百分數時，我們乘以100，因為我們想知道這個小數是100的多少部分。  
如果我們乘以1000或200，我們將得到的是“千分率”或“五分率”，而不是“百分率”。例如，如果我們將0.9375乘以1000，我們會得到937.5，這表示0.9375是1000的937.5部分，這是“千分率”。但在這個問題中，我們需要的是“百分率”，所以我們乘以100。  
希望這個解釋有助於你理解為什麼我們在將小數轉換為百分數時乘以100。如果你還有其他問題或疑惑，請隨時提出。  
16:13:37

所以這一題的答案應該是 93.75% 16:13:42

是的，你完全正確！分數15/16轉換為百分數就是93.75%。你已經成功地將分數轉換為百分數，這正是能力指標5-n-14-S03和5-n-14-S06所要求的。非常好，繼續保持！如果你還有其他問題或需要進一步的解釋，請隨時告訴我。現在我們再來嘗試一題  
如果我的題目是：「 $7/8 = ()$ ， $()$ 的答案應該是幾呢？」  
請你試著解答這個問題，並將你的答案填入空格中。  
16:13:47

$7/8 = 0.875$

教育部因材網 NEW 操作介紹 課程總覽

國立屏東師範大學 學生九年一貫 潘小賢寶祐

訊息5個 待辦0個

我的任務

獎勵

報表

測驗報告

學習紀錄

討論

筆記

提問

討論區

學習扶助

科技化評量

問題回報

© since 2016 國立屏東教育大學 測驗統計與進修學習研究中心

### 對話式機器人

你已經正確地將分數7/8轉換為小數0.875，這是一個很好的開始。但是，題目要求的是將分數轉換為百分數。你記得我們如何將小數轉換為百分數嗎？

a.  $0.875 \times 10$   
b.  $0.875 \times 100$   
c.  $0.875 / 100$   
d.  $0.875 \div 100$

請選擇你認為正確的選項，或者如果你有其他疑問，也可以直接提出。  
16:13:57

B 16:14:02

非常好！你選擇的選項是正確的。我們將小數0.875乘以100來將其轉換為百分數。  
所以，我們的計算會是： $0.875 \times 100 = 87.5$ 。  
接下來，我們在這個數字後面加上百分號(%)，所以我們得到的答案是87.5%。  
所以， $7/8$ 等於87.5%。  
現在，你記得如何將小數轉換為百分數嗎？請隨時告訴我。

Figure 5: The interaction between student and chatbot in the chatbox





Figure 6: the chatbot generating test items to assess students

## 5. 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, its superior accuracy and precision in 5th-grade mathematics, as shown in our initial test, do not guarantee equivalent performance at the high school level. The problem-solving capabilities of GPT-4 would still need to be demonstrated and validated in this more advanced context. 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.

## Acknowledgements

We gratefully acknowledge the National Science and Technology Council of Taiwan, along with the Ministry of Education, for their generous financial support and steadfast endorsement of this study under the Project: MOST 108-2511-H-142 -005 -MY3. 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.

## 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.
- [10] M. J. Gierl and T. M. Haladyna, "Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples," in *Automatic item generation*: Routledge, 2012, pp. 36-49.
- [11] Q. Wu, Q. Zhang, and X. Huang, "Automatic Math Word Problem Generation With Topic-Expression Co-Attention Mechanism and Reinforcement Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1061-1072, 2022.
- [12] J. T. Shen et al., "Mathbert: A pre-trained language model for general nlp tasks in mathematics education," *arXiv preprint arXiv:2106.07340*, 2021.