Automating Academic Assessment: A Large Language Model Approach

2023 7th International Conference on Information Technology (InCIT) | 979-8-3503-5869-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/InCIT60207.2023.10412991

Chatchai Wangwiwattana
University of the Thai Chamber of Commerce
Bangkok, Thailand
Email: chatchai_wan@utcc.ac.th

Yuwaree Tongvivat
University of the Thai Chamber of Commerce
Bangkok, Thailand
Email: yuwaree_ton@utcc.ac.th

***Abstract*—In educational settings, providing timely and quality feedback can significantly enhance student engagement and learning outcomes. However, this task becomes increasingly challenging in larger classrooms. To address this issue, this study introduces a method that leverages Large Language Models (LLMs) — already proven useful in text generation and summarization — for automatically assessing students' short-answer responses. Demonstrating effectiveness across a variety of use-cases such as answer matching, keyword extraction, and clustering, this approach achieves an impressive 99.03% accuracy rate. More than just an automated grading tool, the method also offers the capability to generate tailored, real-time feedback, thus enhancing the efficiency of teachers' evaluation processes. The study further provides suggestions for effectively utilizing LLMs in student assessment tasks.**

Keywords*—*Large Language Models (LLMs), education, grading tools, English learning

I. Introduction

The advanced AI-developed artifacts emerging these days potentially provide ample opportunities to create a large impact on a variety of human-machine interactive activities in a form of written and spoken language. Witnessed in the initiatives of ChatGPT amongst other burgeoning platforms, the tools are able to not only create natural-sounding contents, they can also analyze and evaluate language to a certain extent [1]. These platforms are deployable for education and research in the sense that - with appropriate training - they can create interactive contents, automate personalized learning and grading, and provide tailored feedback, serving as valuable resources to enhance learning outcomes. Integrating these technological tools in education can create a more interactive environment. Meanwhile, with help of machine grading, the amount of time required for evaluating students' responses to tests and providing feedback can be substantially mitigated. Thanks to these advantages, the technology can potentially be useful in large classes and in e-learning platforms.

The widespread adoption of technology in education has become increasingly prominent, serving as a valuable resource for various types of online learning. This trend is clearly evidenced by the popularity of online learning platforms such as MOOCs and HyFlex, which offer students the flexibility to study at their own pace and choose their learning materials, regardless of time or location [2], [3], [4]. These platforms can also complement traditional face-to-face classrooms, offering a more versatile approach to education [5].

While learning autonomy can be fostered through this kind of teaching, the reduced interaction time between students and teachers poses a challenge to effective learning [6]. Furthermore, limited feedback from teachers may curtail students' opportunities for reflection, a critical aspect of the learning process [7], [8]. Teacher feedback is crucial for facilitating students' learning, yet teachers' workload can be substantial, especially in large classes requiring grading of numerous short-answer items.

There is a growing trend to shift evaluation methods toward open-ended questions rather than relying solely on multiple-choice items [9]. Although advancements in online teaching tools have simplified the grading of multiple-choice and straightforward short-answer questions, challenges persist in accurately evaluating responses to open-ended questions within existing Learning Management Systems (LMS).

Previous work has employed TF-IDF and K-means as statistical measures to categorize students' answers effectively, thereby reducing the time teachers need to spend on providing feedback [10]. Despite a satisfactory level of efficiency, understanding the semantic depth of the answers is still required, when automatically-provided feedback is essential. While TF-IDF and K-means have shown promise in categorizing student answers, they lack the capability to fully understand the semantic depth of the content, a crucial element when it comes to automated feedback. This leads us to explore the potential of Large Language Models (LLMs), which have demonstrated advanced language understanding capabilities, for grading tasks to not only improve efficiency but also achieve semantic understanding. With the rapid advancements in LLMs, their effectiveness in language understanding tasks such as text generation and summarization has been proven to be efficient [11]. Yet, their application in automated assessment remains under-explored. In this study, we develop and evaluate a tool that utilizes LLMs for grading tasks, assessing its effectiveness, accuracy, and utility in comparison to traditional teacher grading.

Substantial efforts have been put towards creating effective automatic grading systems in a number of studies [12], [13], [14], [15], [16]. Researchers have explored various approaches to achieve reliable automatic grading, but these often come with computational limitations and concerns about the accuracy of the results. Given that pedagogical designs differ across teachers, students, and institutional cultures, creating a ”fit-for purpose” algorithm for automatic grading would be of advantage. Methods like CBOW and LSTM were previously

979-8-3503-5869-8/23/$31.00 ©2023 IEEE

used to evaluate open-ended answers from reading comprehension [17]. These semantic-based models work well for questions that do not require a specific contextual framework for answers. However, content-based subjects often necessitate specialized or task-specific answers, making both semantics and context crucial for grading. Some other algorithms focus on syntax, employing grammatical structures like POS tagging and parsing to distinguish students' language levels [18].

This study aims to address some of these challenges in evaluating context-based answers. We have developed a tool tailored for grading short, open-ended answers that are specific, unique, and specialized to reading tasks designed by educators. Furthermore, the tool is evaluated for its efficacy in assessing higher-order thinking skills—specifically, how well students understand reading materials and can apply this understanding to new and different contexts. This approach encourages students to actively engage in knowledge creation rather than simply absorbing information from instructional materials [19], [20]. By using LLMs for assisting in grading, the tools help reduce both the time and effort required from teachers as well as opportunity to give personalized real-time feedback to students, thanks to the ability for semantic understanding of LLMs. With this respect, this paper discusses the development of a automatic short answer grading system using LLMs to assess students' responses of three types of short open-ended questions: fixed, multiple choices and open-ended. To understand LLMs further, the effectiveness of the performance of algorithms in the assigned grading and feedback-providing tasks is evaluated.

This paper comprises four main sections, beginning with the introduction in Section I. Section II. describes data collection, data preprocessing and the application. Section III. then describes results and analysis. Finally, Section IV. discusses future research and conclusion.

II. Methodology

A. Data collection

We developed a set of five open-ended questions aimed at assessing students' reading comprehension, which is based on the B1 level of the Common European Framework of Reference for Languages (CEFR). These questions were also designed to encourage critical thinking in accordance with the criteria established by Bloom et al. in 1956[21].

The dataset was designed to evaluate students' reading comprehension. Reading materials for the exercise were sourced from the British Council's LearnEnglish Teens website, specifically the B1-level reading section on food and restaurants. This dataset comprises responses retrieved from 207 students taking an "English for Everyday Communication" course at a university in Bangkok. Five questions were designed to evaluate students' understanding of reading texts and criticality expressed in their responses (Q1-Q5). In keeping up with Bloom et al.'s (1956) criteria [21] for critical thinking skills, questions 1-3 (Q1-Q3) test students' understanding of materials and their ability to apply them to new situations. Students thus needed to demonstrate their 'Understanding' and 'Applying' level of criticality. Q4-Q5 assess students' ability to 'Analyzing' and 'Evaluating' critical thinking skills labelled in Bloom's taxonomy. Thus, these two questions are open to creative responses.

- Q1: Ciaran is a vegetarian. He has a healthy lifestyle and only eats plant-based food. Choose the most suitable restaurant for Ciaran and tell him about your suggestion and why.
- Q2: Lara is new in town. She likes dining at a nice restaurant with a botanical view. She does not eat spicy food. Choose the most suitable restaurant for Lara and tell her about your suggestion and why.
- Q3: Pete is a music lover. He is a big fan of classic rock. He is also a foodie, so he can eat all types of food. Choose the most suitable restaurant for Pete and tell him about about your suggestion and why.
- Q4: Which of these 5 restaurants would you like to visit the most? Why?
- Q5: If you were to open your own restaurant, would you do anything different from these restaurants?

Scores are ranged from 1-3 based on the following criteria. A score of 1 is given for correct answers without reasons provided. A score of 2 is given for answers with some reasons or explanation, and a score of 3 is for answers with reasons and an evaluation for justification (e.g., similar/different/good). For example, answers for Q3 that contain Musical Chairs + Music/Rock/Classical Rock/Classical + Monday were given a score of 3 [10].

B. Data Preprocessing

The purpose of this paper is to develop an automatic grading tool that effectively minimizes the need for teacher intervention in repetitive grading tasks. This in turn could increase the time that teachers may spend on the support for students' learning development in other necessary areas. Subsequently, we evaluated this grading tool by using raw student answers without any prior text processing to examine how well the LLMs can understand the context and assess student responses similar to the way a teacher would, shown in fig 1. A total of 207 student answers were collected and directly fed into the LLMs system, along with a corresponding question, an example of a teacher's correct answer, and the grading criteria.

C. The Tool

We have developed a web-based assessment tool that utilizes Large Language Models (LLMs) for automatic grading, aiming to minimize or eliminate the need for teacher intervention while provide flexibility for teachers. The tool offers the flexibility to import evaluation criteria in both JSON (shown in figure 3) and human-readable formats, accommodating technical and non-technical users alike. Additionally, it supports batch processing, allowing multiple student answers to be loaded and assessed simultaneously. These features combine to make the tool a robust and efficient solution for automating the grading process in various educational settings.

For the Large Language Model selection, we opted for OpenAI's GPT-3.5. The choice was driven not just by its

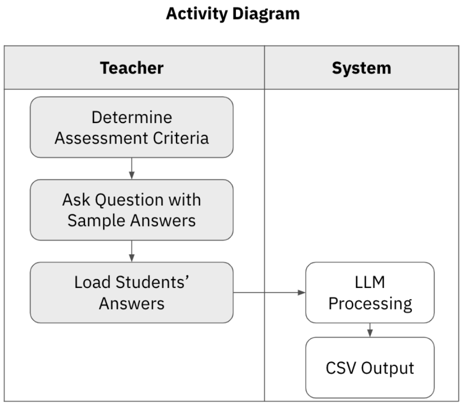
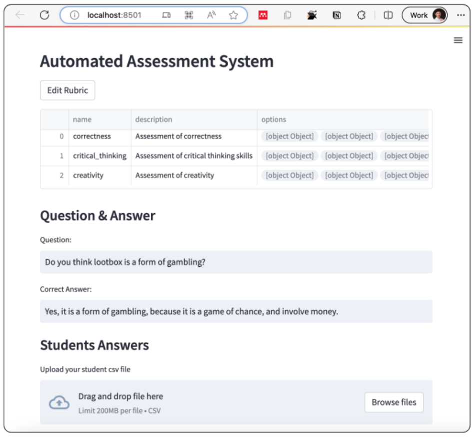


Fig. 1. The process of the system



Automated Assessment System

Edit Rubric

	name	description	options		
0	correctness	Assessment of correctness	[object Object]	[object Object]	[object Objec
1	critical_thinking	Assessment of critical thinking skills	[object Object]	[object Object]	[object Objec
2	creativity	Assessment of creativity	[object Object]	[object Object]	[object Objec

Question & Answer

Question:

Do you think lootbox is a form of gambling?

Correct Answer:

Yes, it is a form of gambling, because it is a game of chance, and involve money.

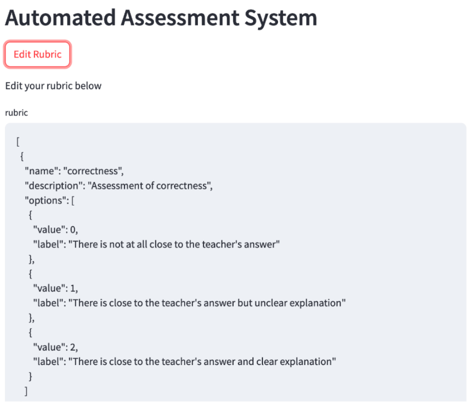
Students Answers

Upload your student csv file

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Fig. 2. The screenshot of the tool interface



Automated Assessment System

Edit Rubric

Edit your rubric below

rubric

```
[
 {
  "name": "correctness",
  "description": "Assessment of correctness",
  "options": [
   {
    "value": 0,
    "label": "There is not at all close to the teacher's answer"
   },
   {
    "value": 1,
    "label": "There is close to the teacher's answer but unclear explanation"
   },
   {
    "value": 2,
    "label": "There is close to the teacher's answer and clear explanation"
   }
  ]
```

Fig. 3. The screenshot of evaluation rubic supported in the tool

exceptional performance but also its accessibility for teachers when compared to other LLM models. Additionally, GPT-3.5 allows developers straightforward access through a simple API. While GPT-4 may offer more powerful capabilities, it is currently slow, expensive, and less accessible[22]. For the scope of this study, GPT-3.5 proves to be sufficient.

Our system gathers necessary information such as evaluation matrix, questions, and example teacher's answer to generate a complete prompt that can be directly fed into the LLM model. The structure of the prompt is divided into five key sections: 1) Role, 2) Question and Answer, 3) Evaluation Criteria and Constraints, 4) Student Responses, and 5) Expected Output. First, we prime the model to assume the role of a teacher's assistant. Second, we provide the question along with the teacher's model answer. Third, the evaluation criteria are supplied in JSON format, enabling easy integration with other systems and a straightforward user interface. Fourth, we include the student responses, which can evaluate multiple students all at once. Finally, we specify the expected output format, asking the model to generate a table with predefined columns namely, StudentID, Score, and Reason. By structuring the prompt in this manner, we guide the model to produce specific content (such as ways of the improvement for each answer) that is useful for both teachers and students. The end result from the system is not only easier to analyze, but also allows the model to deliver additional, pertinent information without going off-topic.

TABLE I. MEAN ABSOLUTE ERROR OF QUESTION 1 TO 3

Question	MAE	STD
Q1	0.5	0.53
Q2	0.49	0.54
Q3	0.42	0.51
Q4	0.49	0.51
Q5	0.61	0.57

III. RESULTS

We examined questions 1-3 by comparing the findings with actual teacher evaluation scores. The mean absolute error was then computed, showing how closely the system evaluation compared to teacher evaluation.

Table I shows mean absolute error(MAE) of each question. The table displays the Mean Absolute Error (MAE) and Standard Deviation (SD) for three questions (Q1, Q2, Q3) when evaluating student performance using a large language model in comparison to teacher evaluations. Given that the rating scale is between 0 and 3, the MAE values ranging from 0.42 to 0.5 are relatively small. This suggests that the evaluations by the large language model are closely aligned with those given by human teachers, especially considering the narrow range of the scale. The SD values are also relatively low, ranging from 0.51 to 0.54. This indicates a relatively consistent level of error across the different questions, suggesting that the model's evaluations are not only close to those of human teachers but also consistently so across different types of questions.

With Q4 and Q5 being more open-ended and aimed at evaluating the students' ability to create new knowledge based on their learning. For Q4, the MAE is 0.49 with an SD of 0.51. These values are closely aligned with the results from Q1 to Q3, suggesting that the large language model performs fairly consistently, even when assessing more open-ended questions. However, for Q5, the MAE is slightly higher at 0.61, and the SD is 0.57. Given that Q5 is more focused on evaluating the students' ability to create new knowledge, this elevated error suggests that the large language model may have limitations in assessing complex cognitive skills like creativity or critical thinking.

TABLE II. EXAMPLES OF SYSTEM FEEDBACK

Student ID	Students answer	Score	Way to Improve
1001	Cafe restaurant with music	1	Add more details about the music theme and how it enhances the cafe experience.
1002	I'm going to make a photo corner in my sweet shop.	1	Consider explaining how the photo corner complements the sweet shop and enhances the overall customer experience.
1003	A restaurant that offers both healthy food and a healthy dessert shop. Because I think everyone will like the dessert way too, but it will be a dessert that doesn't have too much sugar or ingredients.	3	-

One key strength of LLMs is to be able to explain what the model think and reasons to justify the grade they give, as well as give student's feedback to improve the grade to maximize the teacher evaluation criteria in real-time. In order to do that we add additional prompt by using chain of thought in our system as well.

In table II, each row corresponds to a unique student ID, the student's answer to a particular question, the score awarded by the model (on a scale of 0-3), and suggestions for improvement. For example, Student 1001 received a score of 1 for their answer ”Cafe restaurant with music.” The system provides constructive feedback advising that the student ”Add more details about the music theme and how it enhances the cafe experience.” Similarly, Student 1002 scored 1 for proposing a ”photo corner in my sweet shop” and received feedback to elaborate on how the photo corner enhances customer experience. In contrast, Student 1003 scored a 3 for an answer that was already well-developed and detailed, offering a nuanced idea for a restaurant. No additional feedback for improvement was deemed necessary for this student, indicating that their answer already maximized the teacher evaluation criteria.

Qualitative inspection reveals that the system can provide reasonable sound grading to students and offer constructive feedback. This feature proves valuable not only to students but also to teachers, helping them evaluate their own judgment.

Our previous paper purposed an approach for helping teachers to grade students by grouping students result into clusters by using K-mean clustering technique [10]. The result is reasonable well with 99.03% accuracy on grouping restaurants for Q4. In this work, we are using LLMs to extract entities in student answers and group those entity together. In this case, Q4 is ”Which of these 5 restaurants would you like to visit the most? Why?”, we are extracting restaurants, and reasons from each student's answer.

The system isolates two important entities: the name of the restaurant and the reason for the choice. For instance, Student 1001's choice of ”Musical Chairs” and the reason ”because I like music” were successfully extracted. However, in the case of Student 1202, only the restaurant ”Cheesy Bites” was extracted, leaving the reason section empty. In another example, Student 1003's answer lacked any extractable information, resulting in empty fields for both 'restaurants' and 'reasons'. The table also includes a case where the student's answer (from Student 1196) was considered 'Not related to provided context,' demonstrating the model's ability to identify answers that are not aligned with the question's intent.

TABLE III. EXAMPLES OF EXTRACTING IMPORTANT ENTITY

Student ID	Student's answer	restaurants	reasons
1001	Musical Chairs, because I like music.	Musical Chairs	because I like music.
1202	Cheesy Bites	Cheesy Bites	-
1003	The 5th shop is because there is a flower garden.	-	-
1196	Pizza,KFC,Shabu	-	Not related to provided context

The performance is 99.03% accurately classified into group. Taking a closer look at the responses, the model missed in classifying some answers, for example, the score for ”I like to go The Raj the most because are lamb and chicken cooked with mild, medium or hot spices.”. This answer was incorrectly classified. This might be due to that the model became confused after long generating texts. To resolve this issue, we instead tried using a shorter prompt by inputting one student's response. The model can then correctly extract the restaurant and reason. This suggests that longer prompts could confuse the LLM model. Despite the limitation, the performance is sufficient to evaluate student response in real-time.

IV. Discussion

The nature of questions 1-3 is relatively fixed with a certain expectation on the right responses. The outcome demonstrates that a more straightforward method. They system proves works satisfactorily for these specific tasks. The similar MAE and SD values for Q4 and Q5 as compared to Q1 to Q3 suggest that the model performs consistently across multiple types of questions. Nevertheless, the elevated MAE for Q5 hints at limitations when evaluating higher-order cognitive skills. The increased MAE in q5 suggests that evaluating creative skills and the ability to generate new knowledge could be a challenging task for the language model. One of the remarkable strengths of the LLM is its ability to offer real-time feedback that is not only evaluative but also formative. This allows students to immediately understand where they fell short and how they can improve. For more complex cognitive demanding as such question 5, the system employs a 'chain of thought' to ensure the LLM explains its reasoning, thereby increasing the transparency and educational value of the evaluation. As seen in the examples, the LLM's feedback is tailored to the quality of the student's answer. Students with less developed answers receive specific suggestions for an improvement, while high-performing students receive affirmation for their well-thought-out answers. This study focuses on evaluating short answers. The longer essay questions with more complex cognitive evaluation are yet to be further explored.

V. Conclusion

This preliminary study pioneers the use of Large Language Models (LLMs) as an automatic tool aimed at aiding educators in the evaluation process, as well as evaluating the performance of the grading system. This tool can adapt to a variety of educational settings where short-answer evaluation styles are employed. We have explored the use of LLMs for tasks such as adopting entity extraction and providing real-time, formative feedback, complementing our previous efforts by using LLMs for evaluating and clustering student answers. Our work offers several key contributions and avenues for future research:

- Entity Recognition and Reasoning: One of the advancements in using LLMs is their ability to extract and group important entities from student answers. This enables a more nuanced and granular level of assessment compared to traditional clustering techniques.
- Feedback and Chain of Thought: The LLM's capabilities extend beyond grading to include the provision of real-time feedback and explanations for the grades given. This fills an essential gap in the assessment process, making it both evaluative and formative.
- Adaptation to Diverse Learning Environments: The flexibility of the LLMs-based system allows it to adapt to various evaluation styles and criteria, offering a degree of customization not easily achieved with other automatic grading systems.

While promising, the utility of LLMs in various educational contexts and among learners with differing language proficiency levels still warrants further exploration, such as multi model large language models. These have the potential to capture the semantics and context in student answers as well as understand visual and auditory responses all together. This research study serves as a foundational step in leveraging AI technologies to make the evaluation process in educational settings more efficient and insightful.

REFERENCES

[1] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei Openai. Language Models are Few-Shot Learners. Technical report, 2020.

[2] Mariam Mouse Matta Abdelmalak and Julia Lynn Parra. Expanding Learning Opportunities for Graduate Students with HyFlex Course Design. *International Journal of Online Pedagogy and Course Design*, 6(4):19–37, 10 2016.

[3] Brian Beatty. Hybrid courses with flexible participation: The hyflex course design. *Practical Applications and Experiences in K-20 Blended Learning Environments*, pages 153–177, 12 2013.

[4] Michael Detyna, Rodrigo Sanchez-Pizani, Vincent Giampietro, Eleanor J. Dommett, and Kyle Dyer. Hybrid flexible (HyFlex) teaching and learning: climbing the mountain of implementation challenges for synchronous online and face-to-face seminars during a pandemic. *Learning Environments Research*, 4 2022.

[5] Stephan A. Curiskis, Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, 57(2):102034, 3 2020.

[6] Benjamin Luke Moorhouse, Yanna Li, and Steve Walsh. E-Classroom Interactional Competencies: Mediating and Assisting Language Learning During Synchronous Online Lessons:. *RELC Journal*, 2 2021.

[7] Steve Mann and Steve Walsh. Reflective Practice in English Language Teaching: Research-Based Principles and Practices. *Reflective Practice in English Language Teaching: Research-Based Principles and Practices*, pages 1–292, 6 2017.

[8] Steve Walsh and Steve Mann. Doing reflective practice: A data-led way forward. *ELT Journal*, 69(4):351–362, 10 2015.

[9] Yasuhiro Ozuru, Stephen Briner, Christopher A. Kurby, and Danielle S. McNamara. Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, 67(3):215–227, 9 2013.

[10] Chatchai Wangwiwattana and Yuwaree Tongvivat. Semi-automatic short answers clustering and grading with K-Means and Keyword Matching algorithms. In *2022 6th International Conference on Information Technology (InCIT)*, pages 280–284, 2022.

[11] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News Summarization and Evaluation in the Era of GPT-3. 9 2022.

[12] A E E Elalfi, A F Elgamal, and N A Amasha. Automated Essay Scoring using Word2vec and Support Vector Machine. *International Journal of Computer Applications*, 177(25):975–8887, 2019.

[13] Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 2019(8), 2019.

[14] Vivekanandan Kumar and David Boulanger. Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, 5, 10 2020.

[15] Vivekanandan S. Kumar and David Boulanger. Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584, 9 2021.

[16] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495, 3 2022.

[17] Yuwei Huang, Xi Yang, Fuzhen Zhuang, Lishan Zhang, and Shengquan Yu. Automatic Chinese Reading Comprehension Grading by LSTM with Knowledge Adaptation. In *Advances in Knowledge Discovery and Data Mining*, pages 118–129. 2018.

[18] Andrew Kwok Fai Lui, Lap Kei Lee, and Hiu Wai Lau. Automated grading of short literal comprehension questions. *Communications in Computer and Information Science*, 559:251–262, 2015.

[19] Martin Davies and Ronald Barnett. The palgrave handbook of critical thinking in higher education. *The Palgrave Handbook of Critical Thinking in Higher Education*, pages 1–25, 1 2015.

[20] Kate Wilson. Critical reading, critical thinking: Delicate scaffolding in English for Academic Purposes (EAP). *Thinking Skills and Creativity*, 22:256–265, 12 2016.

[21] Benjamin S Bloom, Max D Engelhart, Edward J Furst, and David R Krathwohl. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain.* 1956.

[22] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. 4 2023.