Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation

Zifei FeiFei Han HANZIFEIFEI@GMAIL.COM
Jionghao Lin JIONGHAO@CMU.EDU
Ashish Gurung AGURUNG@ANDREW.CMU.EDU
Danielle R. Thomas DRTHOMAS@CMU.EDU
Eason Chen EASONC13@CMU.EDU
Conrad Borchers CBORCHER@CS.CMU.EDU
Shivang Gupta SHIVANG@CMU.EDU
Kenneth R. Koedinger KOEDINGER@CMU.EDU
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA

Abstract

One-on-one tutoring is an effective instructional method for enhancing learning, yet its efficacy hinges on tutor competencies. Novice math tutors often prioritize content-specific guidance, neglecting aspects such as social-emotional learning. Social-emotional learning promotes equity and inclusion and nurtures relationships with students, which is crucial for holistic student development. Assessing the competencies of tutors accurately and efficiently can drive the development of tailored tutor training programs. However, evaluating novice tutor ability during real-time tutoring remains challenging as it typically requires experts-in-the-loop. To address this challenge, this study harnesses Generative Pre-trained Transformers (GPT), such as GPT-3.5 and GPT-4, to automatically assess tutors' ability of using social-emotional tutoring strategies. Moreover, this study also reports on the financial dimensions and considerations of employing these models in real-time and at scale for automated assessment. Four prompting strategies were assessed: two basic Zero-shot prompt strategies, Tree of Thought prompting, and Retrieval-Augmented Generator (RAG) prompting. The results indicate that RAG prompting demonstrated the most accurate performance (assessed by the level of hallucination and correctness in the generated assessment texts) and the lowest financial costs. These findings inform the development of personalized tutor training interventions to enhance the the educational effectiveness of tutored learning.

Keywords: Large Language Model, Personalized Tutor Training, Automatic Assessment

1. Introduction

The efficacy of one-on-one tutoring for knowledge acquisition and retention continues to gather strong empirical evidence (Kraft and Falken, 2021; Nickow et al., 2020). For tutoring to reach its optimal effectiveness, instructors need a multifaceted skill set, enabling them to provide both content-specific and social-emotional support to students (Thomas et al., 2023; Lin et al., 2023c, 2022b, 2023a, 2022a). However, it is often challenging for novice

© 2024 Z.F. Han, J. Lin, A. Gurung, D.R. Thomas, E. Chen, C. Borchers, S. Gupta & K.R. Koedinger.

math tutors to blend both content-specific and social-emotional support effectively into their teaching. They tend to focus on content-specific instructional guidance, which can result in the inadvertent neglect of social-emotional learning components (Thomas et al., 2022). Social-emotional learning includes indispensable facets such as self-awareness, empathy, adeptness in building relationships, and critical decision-making abilities, which are crucial constituents for fostering the holistic development of students (Jelfs et al., 2009). Prior work shows that neglecting social-emotional learning during tutoring relates to significant loss in tutoring effectiveness (Marshall et al., 2021). To address this issue, assessing social-emotional learning support in tutors can guide personalized tutor training and subsequently improve tutoring effectiveness. However, this assessment has been hitherto expensive and infeasible for many educational applications, as it usually requires human experts, which are scarce (Kraft and Falken, 2021).

In light of recent promising applications of large language models (LLMs) in education (Wang and Demszky, 2023; Dai et al., 2023; Lin et al., 2023b), the present study harnesses Generative Pre-trained Transformers (GPT), such as GPT-3.5 and GPT-4, to automatically assess social-emotional competencies in human tutors (OpenAI, 2023; Lehman et al., 2022). We explore four types of prompting strategies: two basic Zero-shot prompts, Tree of Thought prompting (Yao et al., 2023), and Retrieval-Augmented Generator (RAG) prompting (Lewis et al., 2020). Additionally, given the significance of educational applications at scale, both in research and industry, it is crucial to understand the cost of using GPT models which is vital for assessing the economic feasibility of utilizing the model in tutor training. Consequently, this research investigates two **R**esearch **Q**uestions: **RQ1**: *Can GPT models accurately assess the social-emotional learning competencies of human tutors?* **RQ2**: *How does the performance and cost analysis of GPT-3.5 compare to that of GPT-4 in this context?*

2. Method

2.1. Data

The dataset (tutoring dialogue transcripts) was collected from real-world middle-school math tutoring sessions in the United States. These sessions were conducted on Zoom between September 2023 to October 2023, with novice human tutors to teach math ranging from Grade 6 to Grade 8. It should be noted that prior to the tutoring sessions, these human tutors completed some lessons involved social-emotional learning from a tutor training platform as described in Lin et al. (2023c). Each session, lasting approximately 30 minutes, involved dividing a group of students into multiple breakout rooms. Tutors were responsible for managing three to five breakout rooms, conducting one-on-one tutoring as they circulated among them. When tutors noticed that certain students were capable of working independently with little assistance, they would move on to aid another student in a different room. The entire tutoring process was recorded via Zoom, and the recordings were transcribed using the speech-recognition tool Whisper (Radford et al., 2022). We randomly selected five tutoring transcripts where a math tutor engages in tutoring sessions with five students, guiding them through the process of solving arithmetic problems.

2.2. Assessing Tutoring Practices

The tutoring dialogue transcripts were evaluated based on research-based principles in social-emotional learning and relationship building proposed by Chhabra et al. (2022). The details of these principles are presented in Appendix B; broadly we considered 5 categories: *(1) Giving Effective Praise*, *(2) Supporting a Growth Mindset*, *(3) Reacting to Errors*, *(4) Responding to Negative Self-Talk*, and *(5) Using Motivational Strategies*. We used principles from these five most frequently used tutoring strategies to develop a rubric for assessing the tutoring practice.[1] For brevity, we take the principle of *Giving Effective Praise* as an example. This principle suggests that during tutoring, the instructor should focus on the praise on student learning effort instead of outcome. A desired praise is *"You are almost there! I am proud of how you are persevering through and striving to solve the problem. Keep going!"* while an undesired praise is *"You are so smart and almost got the problem correct."* The principles and rubric from the five lessons further inform the design of prompt strategies to evaluate tutor's use of the practice.

2.3. Large Language Model Generated Evaluation and Feedback

To answer RQ1, we designed four types of prompt strategies. The rationale for each prompt design follows two main elements where the prompt instructs the GPT model to 1) generate an evaluation score and 2) provide evidence from dialogue as rationale for the score. By doing so, we designed four type of prompting strategies (see Appendix A) to analyze the ability of GPT models on evaluating tutor performance on social-emotional learning, and these prompting strategies are detailed below:

Basic Zero-shot Prompt Type I. This zero-shot prompt was designed by using five effective tutoring principles of social-emotional learning and relationship building (detailed in the Appendix A). We prompt the GPT models to assess the whole tutoring transcript based on the tutoring principles in terms of scores and interpretation of score.

Basic Zero-shot Prompt Type II. This zero-shot prompt (detailed in the Appendix A) instructs the model to first identify the incorrect use of tutoring practice from the transcript, and then assess the scores and provide interpretation of the scores.

Tree of Thoughts (ToT) Prompt. As reported by Yao et al. (2023), the Tree of Thought (ToT) prompt allows information to be organized in a structured way, like branches of a tree. This structured format helps the language model understand the information more comprehensively compared to simple and linear prompts. As a result, when using the Tree of Thought prompt, the language model might generate more precise and detailed responses. Our proposed ToT prompt (detailed in the Appendix A) is shown in Figure 1. The GPT models take the tutoring transcript as input. Then, it evaluates the tutor's social-emotional learning based on each principle's rubric (e.g., Using Motivation Strategies in Figure 1). As a result, the model generates scores and interpretation of the scores on assessing the tutor's social-emotional learning competencies in tutoring.

Retrieval-Augmented Generation (RAG). As reported by Lewis et al. (2020), Retrieval-Augmented Generation (RAG) often proves more effective than zero-shot prompts due to its integration of external knowledge sources (e.g., transcriptions and principles of tutoring), enriching the understanding of the GPT model and enabling it to generate more relevant and

1. The details of five lessons can be found via https://www.tutors.plus/solution/training

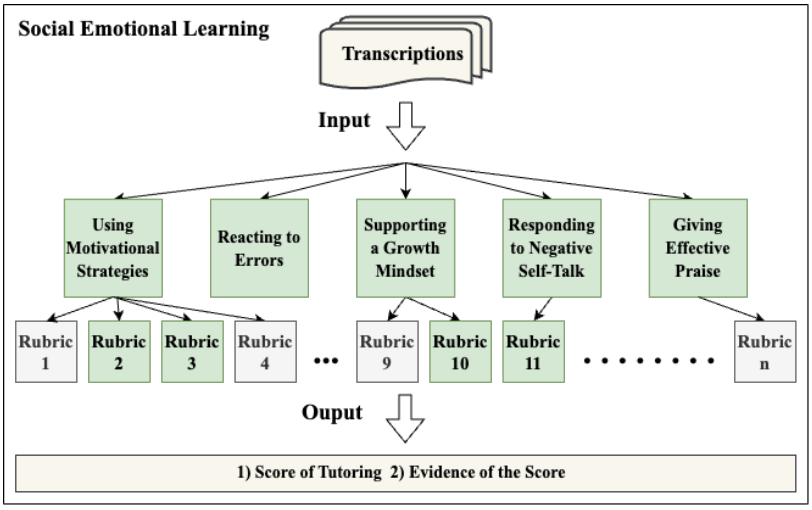

Figure 1: Tree of Thought Evaluation Framework

accurate responses. Our study developed an RAG-based prompt (detailed in the Appendix A) as depicted in Figure 2. Within the information database component (see Figure 2), tutoring transcriptions and principles on social-emotional learning are initially converted into word embeddings (i.e., words represented as vectors preserving semantic information, as described in Kusner et al. (2015)) stored within the database. These embeddings form the basis for our RAG-based prompt, allowing the GPT model to access a broader and more relevant set of information. In the actual evaluation process, the RAG model's evaluation engine (see Figure 2) can selectively retrieve and incorporate information from our prepared word embeddings, guided by the principles of social-emotional learning. This selective retrieval ensures that the GPT model focuses on the most relevant aspects of the tutoring transcripts, ultimately enabling the RAG-based prompt to generate scores and provide evidence for these scores.

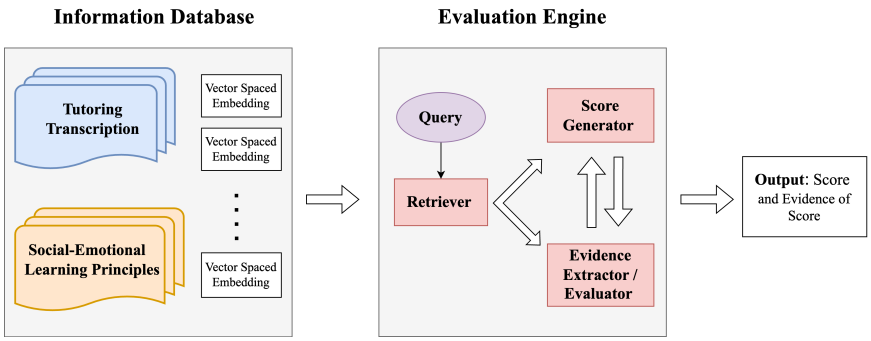

Figure 2: The structure of Retrieval-Augmented Generation (RAG) based prompt

2.4. Evaluation Metrics

Annotation was conducted by a single human coder since we aim to establish a preliminary understanding of evaluating the accuracy of GPT models' output, which can be later expanded and refined in subsequent phases of research. While the use of one coder suits our current exploratory needs, we recognize its limitation (e.g., with respect to inter-rater reliability). The human coder annotated the GPT generated output based on two metrics: 1) **Correctness** and 2) **Hallucination**, which are detailed below:

- **Correctness** is the metric to evaluate the capability of GPT models in accurately assessing the tutor's use of social-emotional learning principles within the tutoring transcript. Evaluating correctness is essential for verifying the model's understanding and interpretation of the given information. Human coders review the GPT-generated assessment results, which include feedback and scores, against the actual tutoring transcript. The scoring categories for this metric include *"-1"* indicating that no information was generated for a specific social-emotional learning principle, *"0"* for incorrect model assessment, and *"1"* for correct model assessment. The correctness is guided by a rubric based on the five principles described in Section. 2.2.

- **Hallucination** is the phenomenon of generative models producing content that deviates from factual accuracy or logical coherence with respect to the input prompt or source content Ji et al. (2023). Hallucinations may manifest as fabricated facts, illogical statements, or irrelevant responses that do not align with the established context or contradict the known data Ji et al. (2023). In our study, we note instances where the GPT model generated feedback that was unrelated to events in the tutoring transcript, exemplifying the issue of hallucination. This issue poses significant challenges for applications that rely on the veracity of generated text, necessitating rigorous validation mechanisms to ensure the reliability of model outputs in critical domains. To measure the hallucination of generated text, the human coder annotated the output into *"-1"*, *"0"*, *"0.5"*, and *"1"* where *"-1"* indicates that no information was generated for a specific social-emotional learning principle; *"0"* indicates no hallucination; *"0.5"* signifies partial hallucination (both hallucinated and non-hallucinated information coexist in the text), and *"1"* represents a completely hallucinated response.

Evaluation on Financial Cost. To answer RQ2, we recorded the cost for using different GPT models with different prompts. We counted the input token from prompt and transcript whereas the output tokens from GPT generated text. We called the API of GPT-3.5 Turbo and GPT-4 Turbo. By referring OpenAI's GPT API pricing as of November 2023,[2] we calculated the cost associated with each generated text from GPT models.

3. Results

Results on RQ1. Figure 3 illustrates the accuracy of GPT models when evaluating the tutoring transcript across different prompt strategies. Accuracy was measured by hallucination and correctness metrics (detailed in Section 2.4), which aimed to assess GPT models'

2. https://openai.com/pricing

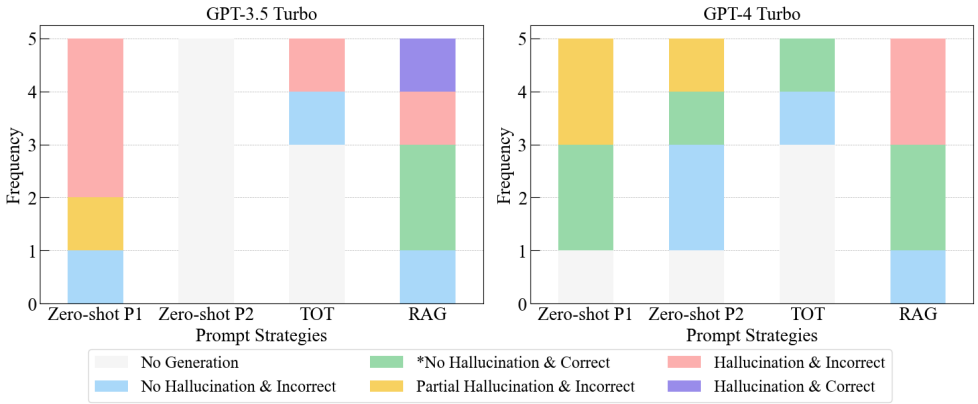

Figure 3: Analysis of GPT models' accuracy across various prompt strategies. Zero-shot P1 and P2 denote the basic Zero-shot prompt type I and type II, respectively

ability to evaluate tutor competency in social-emotional learning. Notably, the green area in Figure 3 represents instances of no hallucination and correct evaluation, which is the desired evaluation on the tutoring transcript. The left side of Figure 3 showed the accuracy of the GPT-3.5 model. This demonstrates that the RAG strategy has the potential in providing no hallucination and correct transcript evaluation. In comparison, other prompting strategies from GPT-3.5 failed to provide the desired transcript evaluation. It should be noted that the results from GPT-3.5 include an evaluation stating, *"The tutor did not respond to negative self-talk by validating the student's feelings or building their self-efficacy."* On closer examination of the tutoring transcript, we observed that the student did not engage in negative self-talk, hence the tutor's lack of response. However, the GPT-3.5 evaluation implied that the tutor intentionally did not respond to negative self-talk, constituting a model hallucination. In contrast, the GPT-4 model results (on the right in Figure 3) show a generally more accurate evaluation (highlighted in green) compared to the GPT-3.5 model, demonstrating its advanced capabilities on this task. It is important to note that both the RAG and Zero-shot (P1) prompts in GPT-4 yielded more accurate evaluations than other prompting strategies. The comparative analysis of GPT-3.5 and GPT-4 underscores that the RAG-based prompting strategy consistently produces the desired output, highlighting its effectiveness across different model versions.

Results on RQ2. In our subsequent analysis, we evaluated the financial cost per lesson evaluation for both GPT-3.5 Turbo and GPT-4 Turbo using different prompting strategies, as outlined in Table 1. The financial cost associated with GPT-4 is approximately 10 times greater than that of GPT-3.5. The Retrieval-Augmented Generation (RAG)-based prompting strategy was identified as the most cost-effective for both the GPT-3.5 and GPT-4 models. Related to the RQ1 results, the RAG-based approach has demonstrated its capability to provide accurate evaluations without hallucinations in tutoring transcripts for both models. Consequently, the results suggest that the RAG-based prompt is the most cost-efficient strategy. These findings provide insights into the selection of different

prompting strategies and GPT models, particularly in terms of balancing effectiveness with financial constraints.

Table 1: Comparison of costs across various prompting strategies in GPT-3.5 and GPT-4

Prompt	GPT-3.5 Turbo	GPT-4 Turbo
Zero-shot Prompt Type I	$0.100	$1.035
Zero-shot Prompt Type II	$0.014	$0.188
Tree of Thoughts (ToT)	$0.013	$0.137
Retrieval-Augmented Generation (RAG)	$0.008	$0.137

4. Conclusion

This study highlights the potential of Retrieval-Augmented Generation (RAG) prompting in evaluating the quality of tutoring based on social-emotional learning competencies. By integrating tutoring transcripts and principles into the GPT model via word embeddings RAG enables more contextually relevant and precise evaluations. The RAG-based prompt not only showcased more accurate performance in evaluating tutoring practices and lower financial costs compared to other prompts but also laid the groundwork for broader applications in tutor skill assessment. Moving forward, we aim to further explore the RAG prompt's effectiveness in assessing additional tutoring competencies, such as building content skills and promoting inclusion. Additionally, we plan to employ the RAG prompt to evaluate a broader range of real-world tutoring transcripts, assessing its efficacy across diverse tutoring interactions. Our goal is to identify areas where tutors may lack skills and provide corresponding training lessons to help them improve. Specifically, we intend to design a training lesson recommender system that can process the assessment from the GPT model and offer lesson recommendations to assist tutors in enhancing their tutoring skills. A prototype developed for the demonstration of the lesson recommendation system is accessible at https://tutorevaluation.vercel.app.

References

Pallavi Chhabra, Danielle Chine, Adetunji Adeniran, Shivang Gupta, and Kenneth Koedinger. An evaluation of perceptions regarding mentor competencies for technology-based personalized learning. In *Society for Information Technology & Teacher Education International Conference*, pages 1620–1625. Association for the Advancement of Computing in Education (AACE), 2022.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

Anne Jelfs, John Richardson, and Linda Price. Student and tutor perceptions of effective tutoring in distance education. *Distance Education*, 30, 11 2009. doi: 10.1080/01587910903236551.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Matthew A Kraft and Grace T Falken. A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, 7:23328584211042858, 2021.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. Evolution through large models, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Jionghao Lin, Mladen Rakovic, David Lang, Dragan Gasevic, and Guanliang Chen. Exploring the politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 282–293, 2022a.

Jionghao Lin, Shaveen Singh, Lele Sha, Wei Tan, David Lang, Dragan Gašević, and Guanliang Chen. Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127:194–207, 2022b.

Jionghao Lin, Mladen Raković, Yuheng Li, Haoran Xie, David Lang, Dragan Gašević, and Guanliang Chen. On the role of politeness in online human–human tutoring. *British Journal of Educational Technology*, 2023a.

Jionghao Lin, Danielle R Thomas, Feifei Han, Shivang Gupta, Wei Tan, Ngoc Dang Nguyen, and Kenneth R Koedinger. Using large language models to provide explanatory feedback to human tutors. 2023b.

Jionghao Lin, Danielle R Thomas, Zifei Han, Wei Tan, Ngoc Dang Nguyen, Shivang Gupta, Erin Gatz, Cindy Tipper, and Kenneth R Koedinger. Personalized learning squared (plus): Doubling math learning through ai-assisted tutoring. 2023c.

Lydia Marshall, Jonah Bury, Robert Wishart, Rebekka Hammelsbeck, and Emily Roberts. The national online tuition pilot. *Education Endowment Foundation. https://educationendowmentfoundation. org. uk/projects-and-evaluation/projects/online-tuition-pilot*, 2021.

Andre Nickow, Philip Oreopoulos, and Vincent Quan. The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. working paper 27476. *National Bureau of Economic Research*, 2020.

OpenAI. Gpt-4 technical report, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

Danielle Thomas, Cassandra Brentley, Carmen Thomas-Browne, J. Richey, Abdulmenaf Gul, Paulo Carvalho, Lee Branstetter, and Kenneth Koedinger. *Educational Equity Through Combined Human-AI Personalization: A Propensity Matching Evaluation*, pages 366–377. 01 2022. ISBN 978-3-031-11643-8. doi: 10.1007/978-3-031-11644-5_30.

Danielle Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth Mclaughlin, and Kenneth Koedinger. When the tutor becomes the student: Design and evaluation of efficient scenario-based lessons for tutors. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 250–261, 2023.

Rose E. Wang and Dorottya Demszky. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Appendix A. Prompting Strategies

Table 2: Basic Zero-shot Prompt Type I

	Prompt
Scoring	*Given a dialogue of a tutoring session, please evaluate the Tutor based on specific best teaching practice within {`Principle_Name`} Please return 1 if the tutor correctly used the tutoring practice. Return 0 if the tutor incorrectly used the tutoring practice. Please only return 0 or 1*
Generator	*Please briefly explain why you give the score?*

Table 3: Basic Zero-shot Prompt Type II

	Prompt
Incorrect Identification	*Given the following evaluation criteria {`Principle_Criteria`} Please identify if there is any tutor's incorrect use of the tutoring strategy {`Principle_Name`}based on the criteria above.If there is incorrect response, return the incorrect response by tutor as evidence in the dialogue and list the criteria not met. If the tutor used teaching strategy correctly, please return based on the criteria above, which ones are correct and also return evidence from the dialogue.*
Score Generation	*Return the score of {`Principle_Name`} from 0 to 5 based on the evaluation. Give one point to each criteria met.*

Table 4: Tree of Thought Prompt

	Prompt
Layer_1	*{`Social_Emotional_Learning_Principles`} For the following transcript between a tutor and a middle school student, score how well the tutor performed in the competency area above. Give one point for each of the following criteria or skills being met by the tutor. For example, if a tutor did not demonstrate any evidence of a given skill or criteria give a score of 0. If a tutor met all the given criteria, give a score of 5. Please only return the evaluated score from 0 to 5.*
Layer_2	*For each criteria listed, please indicate which from the current {`Social_Emotional_Learning_Principles`} is not met, and which criteria are met.*
Layer_3	*Given a dialogue of a tutoring session between a tutor and a middle school student, please evaluate the Tutor based on specific given criteria : {`rubric`}, Please return 1 if the tutor correctly used the tutoring practice. Return 0 if the tutor incorrectly used the tutoring practice. Provide your evaluation in the form of a number. Please also list evidence why you provide the evaluation.*

Table 5: Retrieval-Augmented Generation Based Prompt

	Prompt
Retriever	*For each criteria and rubric above, please identify all of tutor's correct and incorrect use of the practice above. Return the dialogues of the tutor as evidence in the format: 1. Competency 2. Each Criteria of the competency 3. Sentences that tutor said within the dialogue serves as evidence.*
Generator	*Return the score of {`Principle_Name`} from 0 to 5 based on the evaluation. Give one point to each criteria met. Please only return the evaluated score from 0 to 5.*

Appendix B. Social-Emotional Learning Principles

Table 6: Social Emotional Learning Principles

Principles	Description
Giving Effective Praise	Praising students for putting forth effort by giving process-focused praise instead of praising students for getting an answer correct or getting a good grade
Supporting a Growth Mindset	Supporting a growth mindset instead of a fixed mindset by encouraging students on the learning process and not necessarily just getting the answer
Reacting to Errors	Responding to students when students make errors or mistakes, by not directly calling attention to the error but guiding students to realize and correct the error themselves.
Responding to Negative Self-Talk	Responding to students positively when students engage in negative self-talk, such as saying "I can't do this" or "this is too hard for me" by validating a student's feelings but encouraging and building their self-efficacy
Using Motivational Strategies	Rewarding students by using intrinsic and extrinsic motivation strategies, such as rewarding students for working hard by giving them time at the end of a session to discuss their interests