Rubric-Based Assessment of Narrative Texts via Human-AI Collaboration: A Specialized GPT Model Approach

Evaluación de Textos Narrativos Basada en Rúbricas con Colaboración Humano-IA: Enfoque Especializado de Modelo GPT

https://doi.org/10.4438/1988-592X-RE-2025-411-730

Tolga Demir
https://orcid.org/0000-0002-1812-3397
Republic of Türkiye Ministry of National Education

Sait Çüm
https://orcid.org/0000-0002-0428-5088
Dokuz Eylul University

Abstract

This study investigates whether narrative texts can be accurately and stably scored over time and whether effective formative feedback can subsequently be provided for these texts through human-AI collaboration. To this end, two models were employed: the default version of ChatGPT and the Text Assessment Tool (TAT), a GPT model specifically trained through a six-step process for this research purpose. 114 narrative texts were scored three times according to criteria in a rubric by both the specially trained and default models. The agreement levels of the scores given by TAT and default ChatGPT with the actual scores, as well as the stability of these scores over time, were examined. The results indicated that, in contrast to the performance of default ChatGPT, TAT's scores demonstrated high levels of agreement with the actual scores and maintained stability over time across all rubric categories, consistently surpassing the threshold and frequently indicating high reliability. Additionally, it was observed that the majority of the feedback generated by TAT met the criteria for effective feedback. The statistical evidence presented in this study underscores that large language models, when specifically trained, can perform very well in scoring texts using a rubric and providing feedback. This is particularly promising for achieving fairer education, especially in large classes and situations where evaluators are overburdened.

Keywords: educational assessment, human-AI collaboration, GPT training

Resumen

Este estudio analiza si los textos narrativos pueden ser evaluados con precisión, mantener calificaciones estables a lo largo del tiempo, y si es posible proporcionar retroalimentación formativa efectiva para estos textos gracias a la colaboración humano-IA. Para ello, se utilizaron dos modelos: la versión estándar de ChatGPT y la Herramienta de Evaluación de Textos (TAT), un modelo GPT específicamente entrenado mediante un proceso de seis pasos diseñado para esta investigación. Se evaluaron 114 textos narrativos tres veces según los criterios establecidos en una rúbrica, utilizando tanto el modelo estándar como el modelo especialmente entrenado. Se analizaron los niveles de concordancia de las calificaciones otorgadas por TAT y ChatGPT estándar con las calificaciones reales, así como la estabilidad de estas calificaciones a lo largo del tiempo. Los resultados mostraron que, en comparación con el desempeño de ChatGPT estándar, las calificaciones de TAT presentaron altos niveles de concordancia con las calificaciones reales y mantuvieron estabilidad a lo largo del tiempo en todas las categorías de la rúbrica, superando de manera constante el umbral e indicando con frecuencia una alta fiabilidad. Además, se observó que la mayor parte de la retroalimentación generada por TAT cumplía con los criterios de retroalimentación efectiva. La evidencia estadística presentada en este estudio demuestra que los modelos de lenguaje a gran escala, cuando son específicamente entrenados, pueden desempeñarse de manera excelente en la evaluación de textos mediante una rúbrica y en la provisión de retroalimentación formativa. Esto es particularmente alentador para lograr una educación más equitativa, especialmente en clases numerosas y en situaciones donde los evaluadores están sobrecargados.

Palabras clave: evaluación educativa, colaboración humano-IA, entrenamiento de modelos GPT.

Introduction

> *"To shorten our path, we needed a horse. We found a wild one, untamed and strong. We had to tame it, for a wild horse would not serve us. This paper is the story of that taming."*
>
> *Sait Çüm & Tolga Demir*

In educational assessment, question types such as multiple-choice, sentence completion, matching, and true-false are frequently employed in both classroom settings and large-scale examinations, particularly for summative assessment purposes, due to their capacity for objective scoring. However, for formative assessments—which aim to identify and address students' learning gaps, monitor their development, and enhance instructional processes—It is essential to utilize open-ended questions, as well as oral and product-based or process-oriented performances that provide more detailed data to the educator. Such assessment approaches enable a clearer identification of students' learning deficiencies and misconceptions, while also facilitating the measurement of higher-order cognitive skills from a taxonomic perspective. Despite these advantages, the time-consuming process of reading, scoring, and providing feedback on these assessments, especially in large classrooms, often leads to their underutilization by teachers.

Recent revolutionary advancements in artificial intelligence (AI) technology have spurred discussions about the future role and significance of AI in our lives. It is now evident that humans are no longer the sole intelligent actors on our planet, making human-AI collaboration inevitable in contemporary organizations (Kolbjørnsrud, 2024). It is not difficult to predict that human-AI collaboration will continue to reduce costs related to time and labor in various fields.

Received: 09/10/2024 Accepted 12/09/2025

Artificial intelligence

Artificial intelligence (AI) encompasses computerized systems that perform tasks and respond in ways typically associated with human intelligence, such as learning, problem-solving, and goal achievement under uncertain and varying conditions. AI has achieved remarkable progress from early problem-solving in the 1950s to the simulation of human reasoning in the 1960s, from initial mapping projects in the 1970s to the advent of intelligent assistants in the 2000s (Dalton, 2024; Fell Kurban & Şahin, 2024). Within this vast domain, generative AI stands out as a specialized subset focused on creating new content that mimics existing data. Up to the present day, the fields of machine learning and artificial neural networks have significantly advanced, enabling the development of sophisticated generative architectures and deep learning algorithms. Notable examples include generative adversarial networks (GANs), variational autoencoders (VAEs), and transformer-based models (Alto, 2023; Chan & Colloton, 2024; Johannesson, 2024), which are integral to the progress and applications of generative AI.

The release of ChatGPT, a generative AI model, by OpenAI in late 2022, made a significant global impact, garnering widespread attention. This development acted as a driving force, encouraging numerous major technology companies to enter the competitive field of generative AI models (Holmes & Miao, 2023). While strong competitors such as Gemini, DeepSeek and Llama have emerged, ChatGPT continues to maintain a slight lead in terms of popularity and widespread user adoption.

ChatGPT, a pre-trained large language model (LLM), utilizes a transformer-based language architecture, a type of deep neural network highly effective for natural language processing (NLP) tasks. It can understand and generate human-like text based on the input it receives. Trained on a vast amount of data, ChatGPT has learned the patterns, styles, and complexities of human language, making it an exceptional tool for communication. Its capabilities have transformed education by offering dynamic human-like conversations, providing instant information, personalized recommendations, and continuous academic support (Chan & Colloton, 2024; Fell Kurban & Şahin, 2024).

Education and LLMs

The popularity of LLMs such as ChatGPT among both teachers and students necessitates research into the alignment of its capabilities and outputs with expectations or defined objectives. We find it crucial to explore the potential of LLMs in facilitating and supporting tasks that are exhausting and time-consuming for teachers, rather than merely assisting with superficial or straightforward tasks. This is particularly important in ensuring the continuity of critical educational processes, such as formative assessment, in large classrooms or periods of high teacher workload, thus sustaining the quality of education.

When the literature on the integration of LLMs into educational practices, particularly in writing skills and assessment, was reviewed, two studies were identified that highlighted AI's effectiveness in generating reading materials (Fitria, 2023; Xiao et al., 2023). Additionally, three studies aligned with our objectives examined AI's ability to provide feedback or score student essays (Steiss et al., 2024; Wang, 2022; Yavuz et al., 2024). However, a rubric-based training process specifically designed for assessing narrative writing skills was not employed in any of these studies.

On the other hand, the use of Large Language Models (LLMs) in education may produce hallucinatory information, leading to accuracy and reliability issues that can negatively affect student learning and critical thinking skills (Elsayed, 2024). Some studies have shown that LLMs, particularly in feedback processes, may fail to fully comprehend student work and at times provide feedback that is either off-topic or superficial (Venter et al., 2024; Jia et al., 2024). This situation may affect the trust that students and teachers place in AI-generated feedback and could lead to more cautious or even skeptical attitudes toward such outputs (Ziqi, 2024). For precisely these reasons, training a specialized, rubric-based model for a specific purpose within the scope of this study is considered important, as it has the potential to overcome some of the challenges that may arise.

The present study

This study aims to determine whether narrative texts can be accurately scored and whether effective formative feedback can be provided through human-AI collaboration. Additionally, the study compares the scoring accuracy and stability over time of a GPT model, the Test Assessment Tool (TAT), which was trained using many-shot iterative prompting approach, with those of the default ChatGPT.

A significant challenge in this study is the inherent subjectivity in scoring narrative texts, even with a rubric. The study suggests that human-AI collaboration can improve objectivity and stability in scoring. For example, determining what constitutes an "engaging title" involves personal judgment, which AI also struggles with. Instead of simplifying the rubric to minimize subjectivity, this research aims to show how human-AI collaboration can develop reliable solutions in contexts requiring subjective evaluation. If successful, this approach could lead to fairer outcomes, even in large-scale assessments or recruitment processes.

The study also investigates the potential of AI collaboration in providing feedback within formative assessment processes to support student development. It posits that such collaboration can reduce teachers' workload in providing feedback on students' work. The effective feedback criteria used to measure the effectiveness of the feedback (Brookhart, 2008; Burke & Pieterick, 2010; Irons, 2008; Juwah et al., 2004) are detailed in Appendix I. Although the process of effective feedback can involve dialogue and face-to-face interaction, this study focuses exclusively on written feedback due to the nature of the materials used.

The following hypotheses guide the research process and analyses:

- H1: The trained large language model will produce more accurate scores for narrative texts compared to the default ChatGPT.
- H2: The trained large language model will demonstrate greater stability and reproducibility in scoring narrative texts over time compared to the default ChatGPT.
- H3: The trained large language model will provide more effective written feedback that enhances students' narrative writing skills compared to the default ChatGPT.

Received: 09/10/2024 Accepted 12/09/2025

Methodology

The methodology of this study comprises two main phases. The first phase involves training GPT specifically for the research objective, while the second phase evaluates the effectiveness of the trained model by comparing it to the default ChatGPT.

For both phases, a rubric from the Turkish Ministry of National Education's 2024 curriculum was used, encompassing eight categories: "page structure," "title," "text structure," "character," "setting," "plot," "language and style," and "spelling and punctuation." Each category is assessed at three levels (details in Appendix II). TAT was trained using the GPT Builder application of OpenAI to scoring texts and provide feedback based on this rubric.

In this study, human evaluators were not used as benchmarks for AI scoring accuracy due to the risk of their evaluation errors introducing additional bias. Instead, the narrative texts in the dataset were created by researchers according to the rubric, with intentional omissions or errors. To ensure the accuracy of the dataset's intended design, a reliability study was conducted with other experts on a subset of the texts, with the results detailed in this section. Following the reliability study, the pre-determined scores of the texts, referred to as "actual scores," served as the gold standard for comparisons.

GPT Training Process

TAT was subjected to a thorough six-step training process, summarized below, using GPT Builder; examples of the prompts used in each step are provided in Appendix III.

- Goal Setting and Initial Assessment

The GPT model is assigned a specific role. The necessary files for this role (preferably in PDF format) are uploaded to the system, and their comprehensibility is verified. In the context of this study, these files contain the criteria specified in the rubric as well as the criteria for effective feedback.

- Criteria Introduction

A question-and-answer session is conducted with GPT about the evaluation criteria. The session aims to determine how well GPT comprehends the criteria and to identify potential issues it might encounter during the evaluation. The prompts given in response to the answers help to clarify how the evaluation criteria are to be applied.

- Example Analysis

Examples (training data) are provided to GPT for many-shot iterative prompting. For scoring tasks, example sentences are presented for texts that could receive 1, 2, or 3 points. For feedback prompts, examples of sentences illustrating effective feedback are provided.

- Upload Sample Files
- When assessing narrative texts, certain formal features—such as paragraph indentation and title placement—also need to be considered, even though they are not directly related to content. Since it was not possible to effectively illustrate these aspects through standard text-based examples, visual samples were required. Therefore, a set of texts containing intentional formatting errors was created and provided to GPT in PDF format. These files served as reference points, allowing GPT to accurately recognize and evaluate formatting elements as part of the assessment process. Structured Practice

Unlike earlier steps that focus on parts, this step aims to see the whole. To elaborate, while previous steps focus on specific criteria within a rubric or a particular aspect of effective feedback, this step observes how GPT scores an entire text and provides comprehensive feedback.

During the training process, the phases of Example Analysis, Upload Sample Files, and Structured Practice can be iteratively repeated to ensure more accurate responses. In cases where the desired outcomes are not achieved, the process is repeated with new examples and the structured texts are re-evaluated for problematic areas.

- Final Evaluation and Confirmation

In this step, the researcher verifies the topics agreed upon with GPT up to this point. Final adjustments are made to the instructions if necessary, and the uploaded source files are confirmed.

FIGURE I. GPT Training process for rubric-based assessment using GPT Builder

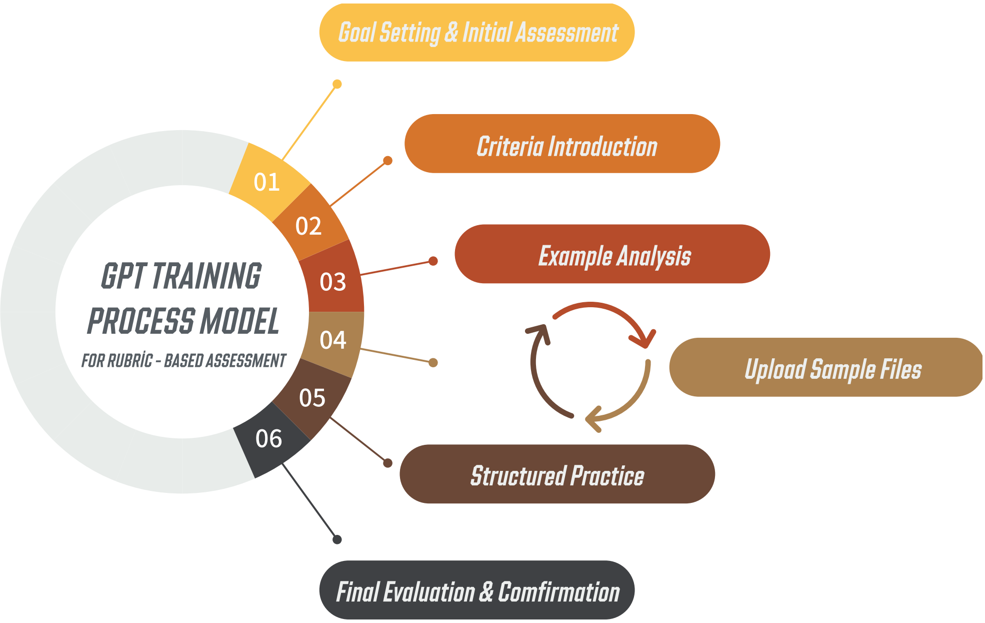

Figure I illustrates the GPT training process, applicable for similar tasks. Click here to access TAT

Challenges and Solutions Encountered During the Training of the AI

In this section, we summarize the notable challenges encountered during the AI training process and the solutions devised, aiming to assist other researchers and practitioners in navigating similar issues.

- File type compatibility

One key challenge was the variety of file types used in the training process. Considering practical applications, evaluators might store texts in different formats, such as photographs of students' written work. We

Received: 09/10/2024 Accepted 12/09/2025

experimented with different file types during the training process to observe any variations in performance. Using .png files resulted in more errors, likely due to the Optical Character Recognition (OCR) process employed by LLM, which made changes to the content before analysis. With .doc files, the model sometimes altered texts, such as adding and scoring titles that weren't originally there. However, using .pdf files minimized these issues, significantly reducing the frequency of such interventions.

- Batch processing vs. individual processing

Another issue involved the mode of text submission—whether collectively or individually. Batch submissions led to significant errors during training and testing, with the model exhibiting unwanted automation in scoring and producing uniform feedback after a few texts. Sequential submission and individual scoring effectively mitigated these issues.

- Text length

Problems also emerged due to the length of the training data. Long texts or prompts in the training set could cause confusion in the trained model. These issues were resolved by organizing training data into shorter, clearer, and more concise segments.

- Connectivity and generalization issues

Occasionally, the trained model struggled to establish the correct connections with previously provided training data, resulting in undesired creativity. This could be due to the model's difficulty in connecting with prior training data, in addition to the challenge of making incorrect generalizations as it is exposed to more varied data. We observed that these connectivity issues resolved themselves over time without new interventions and were not consistently related to specific training data (indicative of randomness). This problem underscores the importance of human-AI collaboration, suggesting that some processes should not be left entirely to AI. Human oversight can effectively eliminate these issues.

Data Collection

114 narrative texts were created for this study, all written in Turkish. These

texts are diverse in terms of evaluation criteria. For instance, some texts consist of a single paragraph but are flawless in terms of grammar and punctuation. Other texts, while ideal in their three-paragraph structure and page layout, lack titles. Some intentionally omit elements of setting. Each story has a unique title, features different characters, and utilizes different elements of setting, resulting in distinct plotlines. In essence, each text is original and unrelated to others. This approach aims to minimize the random factor in scoring or providing effective feedback by TAT.

Narrative texts were uploaded to the Automated Text Analysis Tool (TAT) and scored three times, resulting in scores at three different points in time. The same scoring procedure was applied using the default ChatGPT, yielding three datasets: the actual scores, the scores assigned by TAT across three sessions, and the scores assigned by ChatGPT across three sessions. Additionally, 20 randomly selected narrative texts were re-uploaded to TAT to collect effective feedback based on the previous automated scoring, and outputs were recorded.

Data Analysis

The agreement levels among the actual scores of 114 texts, the scores provided by TAT at three different times, and the scores given by default ChatGPT at three different times were calculated using Krippendorff's α technique. Krippendorff's α is a reliability coefficient commonly used in fields such as social sciences and content analysis to measure the consistency of categorical or continuous data ratings made by multiple raters or coders (Krippendorff, 2004). The following criteria are used to evaluate the obtained Krippendorff's α values: a coefficient value below 0.67 indicates low agreement and reliability, a value between 0.67 and 0.80 indicates acceptable, moderate agreement and reliability, and a value above 0.80 indicates high agreement and reliability.

The effectiveness of the feedback provided by TAT, based on effective feedback criteria, was analyzed using descriptive statistics. In this phase, researchers individually examined a total of 160 feedback for 20 randomly selected texts, using the eight categories in the rubric. The feedback was coded by researchers as successful or unsuccessful according to criteria such as category appropriateness, performance orientation, clarity and comprehensibility, developmental quality, constructiveness, and task specificity. Subsequently, the percentage of successful feedback was calculated relative to the total amount of feedback.

Reliability of Actual Scores

The preliminary scores for 114 narrative texts were determined by researchers using a rubric. To assess the reliability of these scores, a random selection of texts from the 114 was sent to three experts, who provided their opinions on the appropriateness of the scores. The similarity between the expert opinions and the researchers' scores was calculated using the formula $A = C \div (C + a) \times 100$, based on the Miles and Huberman (1994) model. In this formula, A represents the reliability coefficient, C denotes the number of items/terms with agreement, and a denotes the number of items/terms without agreement. According to this model, a similarity ratio of at least 80% is required to achieve consistency. Our reliability study showed agreement rates of 95%, 96.25%, and 91.25% between the preliminary scores assigned by the researchers and the expert evaluations. This consistency justifies considering the preliminary scores assigned by the researchers as the gold standard (actual scores).

Results

All texts in the dataset were scored by TAT at three separate times, and Krippendorff's α was used to determine the level of agreement between each set of scores and the actual scores (t). The results are presented in Table I.

Received: 09/10/2024 Accepted 12/09/2025

TABLE I. Agreement levels for TAT scores with actual scores

Category	α_{t-1}	α_{t-2}	α_{t-3}	α_{mean}	Interpretation
Page structure	0.899	0.897	0.857	0.884	High reliability
Title	0.870	0.802	0.818	0.830	High reliability
Text structure	0.921	0.950	0.950	0.940	High reliability
Character	0.822	0.763	0.807	0.797	Medium reliability
Setting	0.728	0.787	0.759	0.758	Medium reliability
Plot	0.883	0.894	0.901	0.889	High reliability
Language and style	0.874	0.894	0.893	0.887	High reliability
Spelling and punctuation	0.780	0.786	0.814	0.793	Medium reliability

Examining Table I, it is observed that the category with the lowest agreement between TAT's scores and the actual scores is the "setting" category, which involves examining the presence of time and place elements in the stories and their impact. On the other hand, the category with the highest agreement is the "text structure" category, which examines the presence and quality of the introduction, body, and conclusion sections of the stories, with a value of 0.940. Upon reviewing the findings on the agreement levels for each of the three comparisons between TAT scores and actual scores, it was found that all the obtained alpha values, as well as their means, exceeded the threshold considered reliable (0.667).

Along with analyzing the agreement between TAT's scores and the actual scores, the stability of TAT's scores across the three different times was also examined. The results and their interpretations are presented in Table II.

TABLE II. Stability of TAT Scores

Category	$\alpha_{1\text{-}2\text{-}3}$	**Interpretation**
Page structure	0.905	High reliability
Title	0.896	High reliability
Text structure	0.957	High reliability
Character	0.797	Medium reliability
Setting	0.846	High reliability
Plot	0.908	High reliability
Language and style	0.954	High reliability
Spelling and punctuation	0.828	High reliability

Upon examining Table II, it is evident that the scores assigned by TAT exhibit consistency over time across all categories, underscoring the reproducibility of the scoring outcomes. It is observed that the category with the highest agreement between the three different scores assigned by TAT is again the “text structure” category. The “character” category, which assesses the personal and psychological traits of the story characters, shows the lowest agreement. It can be interpreted that the character category, with the lowest alpha value, demonstrates medium reliability, whereas the agreements in the other categories demonstrate high reliability.

Krippendorff’s α values, indicating the agreement between default ChatGPT scores and actual scores, are shown in Table III.

TABLE III. Agreement levels for default ChatGPT scores with actual scores

Category	$\alpha_{t\text{-}1}$	$\alpha_{t\text{-}2}$	$\alpha_{t\text{-}3}$	α_{mean}	**Interpretation**
Page structure	-0.032	0.553	0.370	0.297	Low reliability
Title	0.159	0.422	0.473	0.351	Low reliability
Text structure	0.261	0.521	0.502	0.428	Low reliability
Character	0.266	0.604	0.477	0.449	Low reliability
Setting	0.384	0.469	0.516	0.456	Low reliability
Plot	0.214	0.381	0.477	0.357	Low reliability
Language and style	0.412	0.562	0.550	0.508	Low reliability
Spelling and punctuation	0.233	0.415	-0.171	0.159	Low reliability

Table III shows that “Spelling and Punctuation” has the lowest agreement between ChatGPT and actual scores, while “Language & Style” has the highest. Overall, the default ChatGPT scores show low agreement with actual scores across all categories.

The findings related to the stability of the scores given by ChatGPT, a general language processing model not specifically trained for this research, at three different times are presented in Table IV.

TABLE IV. Stability of default ChatGPT Scores

Category	$\alpha_{1\text{-}2\text{-}3}$	**Interpretation**
Page structure	0.403	Low reliability
Title	0.402	Low reliability
Text structure	0.566	Low reliability
Character	0.475	Low reliability
Setting	0.633	Low reliability
Plot	0.491	Low reliability
Language and style	0.436	Low reliability
Spelling and punctuation	0.627	Low reliability

Examining Table IV, it is observed that the category with the highest agreement between the scores given at three different times is the “setting” category, while the category with the lowest agreement is the “title” category. Interpreting the values in Table 4, it is observed that the agreements for all categories are low reliability.

Received: 09/10/2024 Accepted 12/09/2025

FIGURE II. A comparison between TAT and default ChatGPT concerning agreement with actual scores (left) and the stability of scores (right)

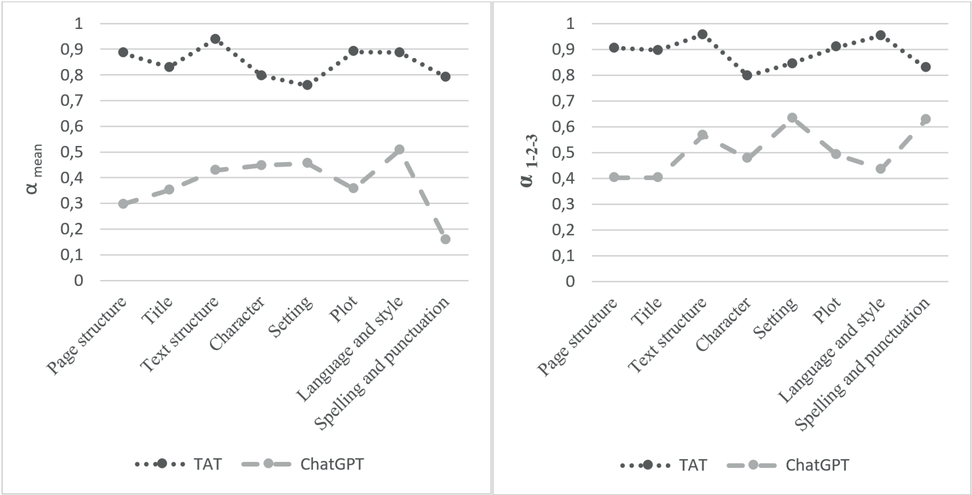

Figure II illustrates the mean alpha values from Tables 1 and 3, as well as the alpha values from the three distinct time points (intra-rater reliability) detailed in Tables 2 and 4. The figure highlights significant discrepancies between the scores given by ChatGPT and those assigned by TAT in terms of both their agreement with actual scores (left) and stability (right)

Based on the evidence regarding the agreement of TAT's rubric-based scoring with the actual scores of the texts and the stability of its scores over different time points, it was concluded that the research hypotheses H1 and H2 were addressed.

To investigate hypothesis H3, the effectiveness of TAT's feedback, given in line with the rubric used in the study, was analysed against the criteria for effective feedback. The feedback provided by TAT was evaluated by researchers using criteria that included being performance-oriented, clear & comprehensible, constructive, developmental, and task-specific, as outlined in the introduction of the study. During these evaluations, it was found that TAT occasionally provided feedback to a category different from the one it should have addressed. To quantify these instances, an additional criterion named "category appropriateness" was defined alongside the effective feedback criteria. The results, including the total number of feedback instances analysed and their success rates, are presented in Table V.

TABLE V. Feedback performance of TAT

Category	Total Feedback	Successful Feedback (%)
Category appropriateness	160	91.88
Performance-oriented	160	100
Clear & comprehensible	160	86.25
Developmental	160	83.75
Constructive	160	100
Task-specific	160	89.38

An examination of Table V reveals that all feedback provided by TAT is performance-oriented and constructive. The criterion with the lowest adherence rate is the developmental criterion, met at 83.75%. Despite the inherent difficulty in crafting feedback that supports student development, TAT's performance in this area is commendably high, as well as high performance being observed across all other criteria.

When feedback is sequentially examined in terms of the criteria that effective feedback should possess, it is observed that 91.88% of the feedback meets the category appropriateness criterion, while 8.12% does not. An example of feedback considered unsuccessful according to this criterion is provided below.

Example Feedback 1: Since the text consists of a single paragraph, the page structure is insufficient. In your next writing, try to use at least three paragraphs, including an introduction, body, and conclusion. By using a different paragraph for each section, you can improve the flow and understandability of your text. For example, in the introduction paragraph, you can describe Can starting to look for his father, in the body paragraph, finding his father and their experiences in the laboratory, and in the conclusion paragraph, provide information about Can's future.

Example Feedback 1 is the feedback provided by TAT for the page structure category of Story 80. However, Example Feedback 1 focuses on suggestions regarding the flow, clarity and elaboration of the story, not the

Received: 09/10/2024 Accepted 12/09/2025

page structure. Although these suggestions are valuable for the development of the text, they do not comply with the rubric criteria for the page structure category. In contrast, Example Feedback 2 is considered suitable for the page structure category.

Example Feedback 2: The text contains two paragraphs that are currently unintended. To enhance the organization and readability of your writing, it is recommended to indent each paragraph in future compositions.

When examining whether the feedback provided by TAT is performance-oriented, it is observed that all 160 pieces of feedback meet this criterion. Example Feedback 3 is an example of feedback provided by TAT that meets this criterion:

Example Feedback 3: The name and some characteristics of the character are mentioned. However, you can add emotional characteristics of the character to help the reader understand the character better and add depth to your story. For example, you can describe Cem's excitement and curiosity during the discovery while progressing through the tunnel. "Cem's heart was pounding as he progressed through the tunnel. He experienced the excitement of a new discovery with every step."

When examining whether the feedback provided by TAT is clear and comprehensible, it is observed that 86.25% of the 160 pieces of feedback are successful according to this criterion. Example Feedback 4 is presented below:

Example Feedback 4: The events progress logically, but you can increase the fluency of the story by adding more details. The story has a clear beginning, development, and conclusion. The preparations for Kaan's party, the fun moments of the party, and his gratitude for the gifts at the end are well described.

Example Feedback 4 is feedback provided by TAT for the plot category for Story 53. This feedback exemplifies clarity and comprehensibility, which are essential principles of effective feedback. It avoids using suggestions that might confuse the student or words that the student might not understand.

An analysis conducted to evaluate the quality of TAT's feedback found that 83.75% of the feedback provided supports the student's development. This is exemplified in Example Feedback 5.

Example Feedback 5: The place and time where the event take place are mentioned, but no detailed information is given. You can strengthen the atmosphere of the story by describing the place and time in more detail. For example, "On a summer day, Mehmet found a silent violin while wandering in the music shop. The shop was filled with old musical instruments."

Example Feedback 5 is feedback provided by TAT for the setting category for Story 15. When this feedback is examined, it is observed that the provided suggestions and examples consist of expressions that support the student's development.

One of the principles of effective feedback is that it should be constructive. According to this criterion, feedback provided to students should encourage them and offer various options instead of rigid commands or instructions. From this perspective, it is observed that all the feedback provided by TAT is delivered in a constructive manner, encouraging the students. This can be seen in the following feedback provided for the setting category for Story 22.

Example Feedback 6: The place and time where the event take place are described in detail. The journey to the library, the ruins, and the manuscripts and books inside the library are clearly described. The contribution of the setting to the story is well emphasized.

One of the qualities that effective feedback should have been that it should be task-specific rather than general. According to this criterion, effective feedback should not use the same expressions for everyone but should be tailored specifically to the student's text. When examining the feedback provided by TAT, it is observed that TAT is quite successful in this regard, with 89.38% of the feedback meeting this criterion.

Example Feedback 7: The events progress logically and in detail. The story has a clear beginning, development, and conclusion. The preparations for Selin's birthday party, the fun moments of the party, and finally, her opening the presents and thanking are well described.

Example Feedback 7 is feedback provided by TAT for the plot category for Story 97. When this feedback is examined, it is observed that it is specific feedback directly related to the text, not general.

Conclusion and Discussion

This study aims to determine if narrative texts can be accurately and stably scored through human-AI collaboration and if effective formative feedback can be provided. Additionally, the performance of the GPT trained for this purpose was compared to ChatGPT, which was not specifically trained for this research, to highlight performance differences.

Conclusion

Scoring accuracy and reliability

Agreement with actual Scores: TAT scored 114 narrative texts using a rubric and the agreement level between the scores and the actual scores for each rubric category was examined. The Krippendorff's α values indicated a strong agreement with actual scores across all criteria, with reliability exceeding the threshold (α ≥ 0.667). The highest agreement was observed in the "Text Structure" category (α = 0.940), while the lowest agreement was in the "Setting" category (α = 0.758).

Stability over time: When examining the agreement levels between TAT's scores at three different times, it was found that Krippendorff's α values were above the threshold (α ≥ 0.667) across all criteria. The scores for the "Text Structure" category demonstrated the highest stability (α = 0.957), while the scores for the "Character" category showed the lowest stability (α = 0.797).

Both in terms of agreement with actual scores and stability, relatively low alpha values were identified in the categories of character, setting, and spelling and punctuation. For the character category, the rubric reveals a subtle distinction between awarding two points and three points. The rubric stipulates that two points should be given when the physical and psychological traits of the characters are described. When these traits, along with the emotions and perspectives that affect the narrative flow, are identified, three points are warranted. Determining which emotion or perspective influences the narrative or distinguishing them can be challenging. This difficulty would challenge a human evaluator as well as TAT's evaluations. In the setting category, the challenge is thought to stem from inconsistencies in the combined portrayal of “place” and “time” elements in the narratives. For example, a story may provide detailed information about the place and its impact on the narrative, but neglect the aspect of time, making it difficult to score according to the rubric, which requires their joint assessment. Further disaggregation of these criteria in the rubric into smaller and clearer components could enhance AI scoring performance. Regarding the spelling and punctuation category, we had to use numerous datasets explaining Turkish spelling and punctuation rules to improve TAT's performance. This necessity is paradoxical because using a large number of datasets can confuse the AI during training. If the stories were in English, fewer datasets would likely have been needed, resulting in better performance. Overall, all performances were above the threshold and satisfactory. Relatively lower performances could be addressed through interventions such as revising the rubric, and these are not viewed as significant issues for text evaluations in AI collaboration

Default ChatGPT's performance: Tests with default ChatGPT revealed findings of low reliability in both agreement with actual scores and internal stability when scoring narrative texts. This was evident even in simple tasks such as evaluating the title of a text. The default model, untrained for text evaluation and unrestricted by specific tasks, often undertakes unwanted tasks such as corrections. For instance, it might add a title to a text that lacks one and then proceed to score the title it added. When examining the scores across different categories, some categories showed very poor performance. For example, the mean alpha value for spelling and punctuation category was 0.159. The default model was particularly weak in examining spelling and punctuation in Turkish texts. This underscores the substantial improvements achieved in initially lower-performing categories following specialized training.

Feedback effectiveness

Criteria compliance: The feedback provided by TAT was evaluated according to the criteria established for effective feedback. The tool demonstrated performance success rates exceeding 83% across all criteria, particularly excelling in delivering performance-oriented, constructive, and task-specific feedback.

Category Appropriateness: Only about 8.12% of the feedback samples were deemed inappropriate for their respective categories, demonstrating TAT's high performance in delivering feedback within the context of each rubric category and effectively reminding students of the relevant criteria. Furthermore, the feedback considered inappropriate was not due to fabricated issues but rather to the confusion between some subtle distinctions among different rubric categories.

Comparative analysis with existing literature

In the study by Yavuz et al. (2024), large language models ChatGPT and Bard were compared for essay evaluation. ChatGPT was used in both its default mode and in a fine-tuned mode with the temperature level reduced to 0.2. The scores given by the AI were compared with those given by human evaluators. The results indicated that both default ChatGPT and fine-tuned ChatGPT, as well as Bard, provided reliable scores. Notably, the fine-tuned ChatGPT showed a very high agreement with human evaluators. In the aforementioned study, the language models were not specifically trained for the task. Fine-tuning was achieved by simply adjusting the temperature setting, which limits the variability of the model's responses. In our study, however, no temperature adjustment was made, and default ChatGPT was used for comparisons. The results of the two studies diverge concerning the performance of default ChatGPT. We considered that the language of the essays being evaluated might be a significant factor. One study used English texts evaluated with an English rubric, while the other used Turkish texts evaluated with a Turkish rubric. To substantiate this claim, more research comparing performances across different languages is required. Another factor contributing to the differing results could be the number of texts evaluated. In the study by Yavuz et al. (2024), only three texts were evaluated, while in our study, 114 texts were evaluated. We observed that as the number of texts to be scored by ChatGPT increased, it produced undesirable automatic responses and applied similar scoring patterns to qualitatively different texts. Thus, the other study may have achieved better performance by evaluating a small number of texts with appropriate prompts and human-AI collaboration. However, we argue that a model specifically trained for a purpose performs much better when there is a heavy lifting to be done.

Awidi (2024), compared human evaluators and default ChatGPT in the evaluation of 108 texts. The intraclass correlation coefficient (ICC) for single measures was 0.349, indicating low agreement, which is consistent with our study's results. Awidi (2024), noted that the agreement increased when looking at average measures and advocated for AI collaboration in text evaluation to achieve more consistent results and significantly reduce human workload.

Regarding the quality of feedback provided to texts, Steiss et al. (2024) compared the feedback quality from humans and ChatGPT on student writings. The study compared 200 pieces of feedback from humans and 200 from AI. The results showed that human raters were more successful in providing high-quality feedback in all categories except for criteria-based feedback. Based on this, the authors argued that ChatGPT can be beneficial in the absence of a well-trained educator. In our study, we achieved quite good results regarding the quality of AI-provided feedback. The difference in results between the two studies is largely due to whether the language model was specifically trained for the purpose. We used a model trained for text evaluation and feedback provision, whereas the other study used a default model. Our study showed that a trained language model excels in delivering effective feedback, which is believed to support student development. Regarding this topic, Escalante et al. (2023) conducted a study to determine how AI feedback and human feedback affect students' writing performance and which type of evaluator the students preferred. The study found no significant difference in performance between the groups receiving AI feedback and those receiving human feedback, and students' preferences for evaluators were evenly split.

Discussion

The results of this study underscore the potential of human-AI collaboration in reliably and objectively scoring narrative texts, even in contexts that require subjective evaluations. The high levels of agreement and stability achieved by TAT, a GPT developed for this study, demonstrate that AI tools, when sufficiently trained, can match human performance in scoring texts and providing effective feedback. The strong potential of AI to support formative assessment processes is particularly significant in densely populated regions and large classrooms, as it can contribute to more consistent and scalable evaluation practices for students while also reducing teachers' workload in monitoring and supporting individual student development. This, in turn, may contribute to a higher quality educational process.

Impact of AI training

The study emphasizes the need for specialized training to improve AI models' proficiency in specific tasks. While ChatGPT excels in general language processing, targeted training is crucial for tasks like evaluating narrative texts. Without task-specific constraints, ChatGPT can produce inconsistent results, which is problematic for both scientific research and practical applications. Thus, the authors advise against using default ChatGPT for critical tasks and recommend employing a trained model with demonstrated reliability.

Not just AI, but human-AI collaboration

The statistical strength of the results produced by the AI in this study provides significant evidence for its use. However, during the process of both training and utilizing the AI, we discovered that it could make unexpected errors in unforeseen areas.

Beyond the difficulties inherent to the task and the influence of subjective decision-making in narrative text assessment, certain deviations in the agreement rates and temporal performance of both the default ChatGPT and TAT can be explained by the phenomenon of hallucination. Therefore, we argue that a completely AI-driven assessment process, devoid of human oversight, would be highly inappropriate. Beyond preventing errors, human-AI collaboration is essential for developing a system that can continually improve and effectively address varying tasks. Periodically feeding the model with appropriate data can greatly enhance its performance and make it more capable of handling diverse situations.

Limitations and recommendations for future research

In the present study, narrative texts were purposefully constructed by the researchers, strictly adhering to a predefined rubric, with intentional incorporation of specific omissions, inaccuracies, and predetermined scoring criteria. This methodological approach enabled a controlled evaluation of the model's proficiency in interpreting and applying evaluation standards. Nonetheless, this design choice introduces inherent limitations. Primarily, the absence of human evaluators and reliance on artificially generated texts may constrain the authenticity and variability that typically characterize genuine student compositions. Consequently, the results obtained from this method may not fully represent the model's potential performance in authentic, real-world educational contexts.

Relatedly, the dataset comprised 114 standardized texts, which, although promoting controlled conditions, might inadequately reflect the diverse range of student profiles and varying writing competencies encountered within large-scale educational environments. To address these constraints, subsequent research could benefit from integrating authentic texts produced by actual students and involving human evaluators to comparatively analyze scoring alignment and temporal consistency of customized GPT models, such as TAT. Furthermore, expanding both the sample size and the dataset diversity might enhance the assessment of the model's generalizability and practical applicability.

Additionally, variations observed between this study and others underscore the importance of investigating how AI language model performance differs across languages. Thus, initiating further practical and experimental research in this area would be beneficial.

Bibliographic references

Alto, V. (2023). *Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation*. Packt Publishing.

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence, 6*, 100226. https://doi.org/10.1016/j.caeai.2024.100226

Brookhart, S. M. (2008). *How to give effective feedback to your students*. ASCD.

Burke, D., & Pieterick, J. (2010). *Giving students effective written feedback*. Open University Press.

Chan, C. K. Y., & Colloton, T. (2024). *Generative AI in higher education: The ChatGPT effect*. Routledge.

Dalton, G. (2024). *Artificial intelligence: Background, risks and policies*. Nova Science Publishers.

Elsayed, H. (2024). The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in AI-assisted education. *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity, 9*(8), 11–23.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(57). https://doi.org/10.1186/s41239-023-00425-2

Fell Kurban, C., & Şahin, M. (2024). *The impact of ChatGPT on higher education*. Emerald Publishing.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum, 12*(1), 44-58. https://doi.org/10.15294/elt.v12i1.64069

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback*. Routledge.

Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024).

Received: 09/10/2024 Accepted 12/09/2025

On assessing the faithfulness of LLM-generated feedback on student assignments. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491–499). https://doi.org/10.5281/zenodo.12729868

Johannesson, P. (2024). *Writing your thesis with ChatGPT: Research, scholarship and academic writing in the age of generative AI*. Kindle Direct Publishing. https://writingyourthesiswithchatgpt.wordpress.com/

Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). *Enhancing student learning through effective formative feedback*. The Higher Education Academy.

Kolbjørnsrud, V. (2024). Designing the intelligent organization: Six principles for human-AI collaboration. *California Management Review, 66*(2), 44–64.

Krippendorff, K. H. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Turkish Ministry of National Education. (2024). *Ortaokul Türkçe dersi öğretim programı*. Millî Eğitim Bakanlığı.

Venter, J., Coetzee, S. A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education, 50*(4), 516–536. https://doi.org/10.1080/02602938.2024.2415649

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech, 40*(1), 80–97. https://doi.org/10.1108/LHT-05-2020-0113

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*

(BEA 2023) (pp. 610–625). Association for Computational Linguistics.

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology,* 56, 150–166. https://doi.org/10.1111/bjet.13494

Ziqi, C., Xinhua, Z., Qi, L., & Wei, W. (2024). L2 students' barriers in engaging with form and content-focused AI-generated feedback in revising their compositions. *Computer Assisted Language Learning*, 1–21. https://doi.org/10.1080/09588221.2024.2422478

Contact address: Dokuz Eylul University, Faculty of Education, Department of Educational Measurement and Evaluation, Izmir, Turkey. E-mail: sait.cum@deu.edu.tr

Received: 09/10/2024 Accepted 12/09/2025

APPENDIX I. The characteristics of effective and ineffective feedback

Category	**Effective Feedback Characteristics**	**Ineffective Feedback Characteristics**
Performance-oriented	1. Feedback is directed towards the performance itself, not the performer. 2. Feedback focuses on specific aspects of the performance rather than general comments.	1. Contains biases towards the student and includes statements targeting their personality. 2. Uses general comments that are not specific to the performance.
Clarity & Comprehensibility	1. Feedback is expressed using words and sentence structures appropriate for the student's age group or developmental level. 2. It clearly specifies what is expected and what constitutes a good performance. 3. Feedback should be detailed and explanatory enough to avoid causing confusion for students.	1. Contains technical and complex expressions that make it difficult for students to understand. 2. Uses vague statements like 'you can do better' instead of specifying what is expected. 3. Feedback is superficial and random, making it unclear what is expected from the student.
Developmental	1. Feedback should include suggestions to help students address deficiencies and achieve the expected performance. 2. Similar tasks or strategies that can be used by the student to facilitate self-learning may be recommended. 3. Emphasize what the student should do first to improve subsequent performances.	1. Emphasizes deficiencies and inadequacies without suggesting ways to address them.
Constructive	1. Feedback should highlight strengths as well as weaknesses in the performance. Good performances should receive feedback as well as poor ones. 2. Use language that encourages the student and supports their self-esteem. 3. Provide options for the student on what they can do, rather than strict commands or instructions	1. Uses patronizing language and statements that passive the student. 2. Includes judgmental or threatening expressions that discourage the student.
Task-specific	1. Feedback should not contain general statements; instead, it should highlight specific points in the student's work and be given specifically in relation to its content.	1. The feedback contains generic statements that could be used for all similar tasks, making the feedback for different tasks appear repetitive and formulaic.

Received: 09/10/2024 Accepted 12/09/2025

APPENDIX II. Analytical rubric for assessing narrative texts (Turkish Ministry of National Education, 2024)

Category	1 Points	2 Points	3 Points
Page structure	The text is not written in paragraphs and is visually disorganized on the page.	The text is written in paragraphs, but the indentations and/or line endings are not properly aligned.	The text is written in paragraphs with proper indentations and line endings, creating a visually organized page.
Title	The text does not have a title.	The text has a title, but it either does not reflect the content or is a common cliché.	The title is relevant to the topic, reflects the content, and is engaging.
Text structure	The text is missing one or more key sections: introduction, climax, and resolution.	The text has an introduction, climax, and resolution, but the transitions between sections are disjointed.	The text has an introduction, climax, and resolution, with logical relationships and smooth transitions between the sections.
Character	Characters are only mentioned by name without any additional information.	Characters are named, and their physical and/or psychological traits are described.	Characters are named, their physical and psychological traits are described, and their emotions, perspectives, and attitudes, which influence the story's flow, are explained or suggested.
Setting	Either the place or the time element is missing or unclear.	The place and time are mentioned, but no detailed information is provided.	The place is well described with auditory and visual details, and the time is detailed, indicating its impact on other story elements.
Plot	There is no clear plot.	There is a clear plot, but the transitions between events are disjointed.	There is a clear plot with strong transitions between events.
Language and style	Most sentences are unclear, lacking semantic and grammatical connections, and the story uses very limited vocabulary.	Sentences are clear and understandable, with some semantic and grammatical connections, but the story uses limited vocabulary.	Sentences are clear and understandable, with well-made semantic and grammatical connections, and the story uses rich vocabulary.

Spelling and punctuation	There are 11 or more spelling and punctuation errors in the text.	There are 6-10 spelling and punctuation errors in the text.	There are no more than 5 spelling and punctuation errors in the text.

APPENDIX III. Example Prompts Used in the GPT Training Process

Steps	Example Prompts
Goal Setting and Initial Assessment	As a language teacher, you will evaluate your students' narrative texts and provide them with effective feedback to help them improve. To do this, you will use the rubric and effective feedback principles documents that I will upload for you.
Criteria Introduction	We will discuss the category of setting. When you examine the criteria, do you see any item on which you might have difficulty deciding? Where do you think you might encounter problems while scoring?
Example Analysis	If the event in the story takes place in the summer, the time is clear; however, if there is no information about the specific details of the time, you should assign a score of 2. The same principle applies to the place element. If the event occurs at an inn and this is mentioned, but there are no detailed descriptions about the inn, you should also assign a score of 2.
Upload Sample Files	Paragraph indentation is when the first line of a paragraph starts further in than the other lines. Now, I will upload a single-paragraph story example without indentation for you. You can use this file as a basis for your evaluation.
Structured Practice	I am going to upload a text for you. Based on our discussions, I would like you to evaluate all sections of the rubric for this text.
Final Evaluation and Confirmation	Now, describe the files I have uploaded to you, summarize the decisions we have made, and specify the rules you will pay attention to during the evaluation.

Received: 09/10/2024 Accepted 12/09/2025