education sciences

MDPI

Article

Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course

Miada Almasre

Information Technology Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; malmasre@kau.edu.sa

Abstract: The recent advancements in the fields of AI technology, generative AI, and Large Language Models (LLMs) have increased the potential of the deployment of such tools in educational environments, especially in contexts where student assessment fairness, quality, and automation are a priority. This study introduces an AI-enhanced evaluation tool that utilizes OpenAI's GPT-4 and the recently released custom GPT feature to evaluate the typography designs of 25 students enrolled in the Visual Media diploma offered by King Abdulaziz University. A mixed methods approach is adopted to evaluate the performance of this tool against the rubric-based evaluations offered by two human evaluators, considering both grading and text feedback. The results indicate that there are statistically significant differences between the AI tool's grading and feedback when compared to that of Evaluator 2; however, none is reported with Evaluator 1. The study presents a qualitative interpretation of the comprehensive feedback by the evaluator and reflects in further research in this area.

Keywords: assessment; ChatGPT; GPT-4; artificial intelligence; typography design; teaching and learning assessment

check for updates

Citation: Almasre, M. Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. *Educ. Sci.* **2024**, *14*, 148. https://doi.org/10.3390/educsci14020148

Academic Editors: Myint Swe Khine and Ernest Afari

Received: 26 November 2023
Revised: 26 January 2024
Accepted: 27 January 2024
Published: 31 January 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The continuous development in the domains of Artificial Intelligence (AI) has impacted various sectors, such as business and healthcare, amongst others [1]. This transformative technology has its imprint in the field of education as well, with a special focus on educational policy, task automation, content and assessment personalization, teacher training, and the practices leading to the improvement of class performance [2]. The optimization AI solutions in the educational field have been augmented recently with the introduction of generative AI models that utilize the capabilities of Large Language Models (LLMs), such as OpenAI's ChatGPT and GPT-4 and Google's Bard [3]. OpenAI's conversational AI technology is basically dependent on the Generative Pre-trained Transformer (GPT) series, with GPT-4 being the latest version, ChatGPT the version of GPT-4, developed by OpenAI, is based in San Francisco, CA, USA. These AI language models have become an important resource and reference points for both educators and students since their release in November 2022. Practices ranging from content creation, assessment design, assessment evaluation, and many others have been optimized using ChatGPT by educators who want to improve their academic performance.

The capabilities of these technologies go beyond text generation and assessment to multimodal generation and evaluation with the introduction of DALL-E 2, an OpenAI model that can generate images and integrate with GPT-4 to offer a critique of image-based content [4]. Other plugins turn GPT-4 into an interface that can read videos and audio as well as offer summarization or even interpretations of them.

There are numerous subject areas in education that rely on visual content creation by educators, as well as assessment types that focus on students' creation/critique of visual content based on a set of specific criteria. Amongst these programs are ones teaching visual

media topics and relevant skills in the fields of typography, design, visual communication, etc. Research on graphic, typographical, emotional, and psychological aspects of visual media in relation to learning experiences and students has been carried out, emphasizing both the relevance and uniqueness of the learning experience from the perspective of educators and students [5,6]. Some of these studies even utilized AI solutions [7].

The assessment component of such subject areas is mostly dependent on scoring learners' designs, which introduce their internalization of all of the theories they have been taught.

Typography courses, in particular, demand the teaching of technical as well as visual communication skills. Learners in such courses should manage the use of types in an artful manner to make textual content comprehensible, readable, and attractive on display. The inherent role of typographical aspects in students' designs then becomes not just aesthetic but informative and communicative. Therefore, assessing students' typographical artwork in an educational context requires that the educator creates a balance between subjective evaluation and objective technical and criteria-based assessment. It is in such contexts that AI, including solutions like ChatGPT and GPT-4, can be used to provide comprehensive and objective evaluations of students' visual designs.

Generally, human educators in various domains are capable of offering learners invaluable insights and an in-depth assessment of their work. However, the persistently implemented traditional methods lead mostly to unsatisfactory results [8]. Amongst the challenges of traditional, human-based evaluations is their inherent subjectivity, which results in inconsistencies and biases. These potential non-uniform evaluation criteria lead to different assessments of students' work for one project, who might eventually become confused about the legibility and clarity of the standards being implemented [9].

In Typography or generally visual media courses, the diverse and complex nature of students' products further complicates the assessment process from the perspective of the educator. Each student's artwork or design represents their individual creative mindset and the skillset they acquired in the course. This diversity demands a compressive and nuanced assessment approach, where each student product is evaluated based on its own merit, as well as considering the set of criteria developed for the course. Creating this balanced assessment approach is always challenging for educators.

Moreover, educators' feedback provision might not be comprehensive in a way that guides the student toward achieving success. Considering the time and workload limitations, educators might not provide learners with detailed or personalized feedback on their design. This challenge demands the development of more efficient methods for automating assessment that can support the educators' expertise [9].

Therefore, the present study aims to utilize the custom GPT feature to design a domain-specific assessment GPT "Typography Evaluator", which is used to grade and provide feedback on typographical designs submitted by students enrolled in a Typography course that is a requirement in the Visual Media Diploma offered at King Abdulaziz University.

The research aims to compare human evaluators' and the Typography Evaluator's output reports of grades and feedback.

This objective is further broken down into three research questions:

1. Is there a significant difference in grading patterns between AI-based and human evaluations of student design projects?
2. Do AI and human evaluators differ in their assessment of specific design aspects such as typeface, format, and color?
3. How do qualitative feedback patterns from AI-based and human evaluators differ when assessing student design projects?

The contribution of this study lies in its attempt to use the recently released (November 2023) GPT-4 feature "custom GPTs" to offer an AI-enhanced assessment solution of visual media produced by students based on domain-specific knowledge for assessment task automation.

The utilization of the custom GPT feature is motivated by the assumption that it will help ensure consistency, objectivity, and efficiency of assessment and feedback. A custom GPT tool is not affected by subjective elements or biases, making sure that all learners receive fair feedback. By depending on predefined rubrics, a custom GPT is capable of providing an impartial evaluation. In addition, in courses with large student numbers and assessments, a custom GPT can accelerate the grading process, saving instructors time. This facilitates teachers to allocate their time and effort in a strategic way, focusing more on the provision of personalized support.

Hence, the rest of this paper is divided as follows: Section 2 presents related work in the context of GPT, assessment, and generally educational research; Section 3 develops the methodology of this research, including participant selection, data collection, tool design, and evaluation processes; Section 4 discusses the findings related to the three research questions; finally, Section 5 provides a complete analysis of the findings. The conclusion offers an overall perspective as well as implications for further research in the field.

2. Related Work

The recent advancements in the field of Generative Pre-trained Transformers (GPT) have impacted various domains, including education. ChatGPT and GPT-4, for example, are not just proficient in NLP tasks but also demonstrate high performance in multimodal understanding, generating text based on text prompts that might take visual information into account. This capability is especially significant in educational assessment, where grading and offering feedback on learners' artwork is mostly subjective. Research in this area is scarce, and this is why, in this review of related research, one must focus on examining contributions from two perspectives: The potential of ChatGPT and GPT-4 in education considering their NLP and the impact of these models on assessment practices.

2.1. ChatGPT and Education

Primarily, OpenAI's ChatGPT is a type of LLM. Like all LLM solutions, it implements deep learning techniques to process and generate text. In this context, ChatGPT has been developed to mimic the human creation of text based on a huge corpus of textual data and is trained to perform different tasks like answering questions, writing, creating code, and generating content. The ChatGPT conversational agent is founded on GPT-3 and GPT-3 Turbo, which has been recently upgraded to GPT-4. The year 2023 witnessed the introduction of plugins as well to GPT-4, the introduction of the DALL-E model, and finally, the release of the custom GPT feature, which allows users to integrate GPT-4, DALL-E and code interpretation and data analysis as enhancements to the custom GPT owned by a community member [3,10].

As [11] reports, ChatGPT and GPT-4 offer users varied features and capabilities that can be used in various domains, such as language understanding and generation through interactive human-AI dialogues. It is understandable that the earliest releases of these models have limited proficiency in domain-specific knowledge; however, allowing users to train their custom GPTs expands the potential of these AI solutions. Considering these models' reasoning skills, they demonstrate great deductive reasoning outcomes but show mixed ones in abductive and inductive reasoning. As far as areas of applications are concerned, ChatGPT and GPT-4 have been used in many fields, including physics, chemistry, medicine, and computer science, indicating varying levels of success [12,13]. Generally, they have showcased capabilities that range from student-level proficiency to specific challenges, particularly in areas requiring specialized knowledge or complex problem-solving with some reported limitations [14].

Research investigating the potential of ChatGPT and GPT-4 in education has started to suggest, for example, frameworks of implementation in various subject areas [15]. For instance, they suggested a framework named "IDEE" for using models like those of OpenAI in education, including tasks such as defining the learning outcomes, setting up criteria of optimized automation, ensuring ethical practice, and assessing the effectiveness of the

academic offering. The researchers even suggested utilizing these models to personalize the learning experience for students.

2.2. Multimodal AI in Instruction and Assessment

Research investigating how multimodal AI systems or applications are used in education for instructional and assessment purposes emphasizes the contributions of such solutions and introduces frameworks for utilizing such systems' data to improve the educational process. However, with the recent developments in Generative AI with applications like OpenAI's ChatGPT and Google's Bard, empirical research discussing the integrated multimodality of these applications and how useful they would be for students' evaluations is non-existent.

Generally, AI solutions' potential in education is explored in [16] with a focus on instructional and assessment tasks. Through a comprehensive survey or related research, the paper advances the argument that AI can positively change education in ways that emphasize the human aspect of it by allowing for novel forms of knowledge representation, feedback, and learning practices and outputs. The research findings demonstrate that AI can be used for formative student assessment, increasing complex epistemic achievement by learners, object-oriented assessment, and multimodal learning. According to this research, multimodal AI solutions can be leveraged to assess students by combining and analyzing data from different modalities, such as text–image or text–audio. The assumption is that such AI models can provide accurate and contextually related responses and facilitate access to applications such as image description, video understanding, and speech-to-text with visual context.

From a technical perspective, there has been research that discussed the potential of recent AI models in the task of interpreting or providing feedback on images. [17], for example, evaluates the accuracy and consistency of three multimodal AI models (Bing's Chat, OpenAI's GPT-4, and Google's Bard) in describing and explaining electrocardiogram (ECG) images. The researchers used a dataset of ECG images from American Heart Association exams and compared the AIs' responses to the human-generated answers. The findings indicate that GPT-4 performed better than other models. Though this research might not be relative to the field of educational assessment, it sheds light on how the multimodal features of these technologies might be implemented to extract feedback on the content of various formats.

2.3. Assessment Evaluation via ChatGPT/GPT-4

Considering assessment in academic settings, ChatGPT can be used to function as an auditor or a control checkpoint of practices related to equity, diversity, and inclusion in educational design, delivery, and assessment, especially areas of the academic process that might be under the influence of biased perspectives [18]. Instructors, as the researchers suggest, can automate, at least partially, the grading and feedback process of student work by using such AI tools to highlight strengths and weaknesses based on set criteria. These kinds of AI integration in the assessment process can save time and enhance outcomes [19].

Research that investigated the utilization of ChatGPT and GPT-4 in the context of students' writing evaluation has focused on grading such written communication using predefined scoring metrics or the potential of providing constructive feedback to students. For example, [20] explores the potential of the GPT-3 text-davinci-003 model as an automated essay scoring (AES) mechanism. The study used 12,100 essays from the ETS Corpus of Non-Native Written English (TOEFL11) as the writing dataset, which was scored using ChatGPT. The model was prompted using the IELTS TASK 2 writing band descriptors as a rubric. Then, the reliability and accuracy of the scores were assessed against established benchmarks. The results indicated that the role of linguistic features in enhancing the accuracy of AES using GPT is very important and reported that the implemented process has achieved a certain level of accuracy and reliability, which makes it a potential tool in the hands of educators.

Amongst the studies that evaluated GPT-4 for its ability to score the coherence of students' written work is [21]. This paper utilizes GPT-4 to perform an automated writing evaluation (AWE) process, which focuses on rating the discourse coherence of the written texts. The research adopts a methodology by which a dataset consisting of 500 test-taker written texts from the Duolingo English Test, which are scored for coherence by human experts on a 6-point scale, is re-rated using GPT-4. Basically, three prompt configurations were used: rating then rationale, rationale then rating, and rating only. Findings report that GPT-4 ratings demonstrate a high agreement with human ratings, scoring a QWK of 0.81 and a Spearman correlation of 0.82. Also, GPT-4, as this paper concludes, has the capability to assess discourse coherence and provide scores that are comparable to educators' ratings.

Similarly, [22] examined the use case of utilizing ChatGPT to generate written feedback on learners' research proposals in a data science course. To assess the quality of the generated feedback, the research evaluated the readability, agreement, and feedback effectiveness. Readability is a Likert-based measurement (on a 5-point scale) of generated feedback. Agreement evaluates, using five assessment metrics, the consistency between ChatGPT and teachers' feedback, considering a polarity of positive vs. negative. To evaluate feedback effectiveness, the researchers considered the presence of the four dimensions of feedback: task, process, regulation, and self, which were presented by Hattie and Timperley (2007). Generally, the findings demonstrate that ChatGPT potentially generates more readable, detailed, and fluent feedback than the teacher. However, ChatGPT scored a high agreement level with the instructor regarding the topic of assessment. Moreover, it is reported that ChatGPT can generate process-related feedback more than task-related feedback.

In addition, [23] focuses on investigating the linguistic and instructional quality of student-generated short-answer assessment questions. The researchers implemented a machine learning model, human experts, and GPT-3 automated feedback to evaluate the questions generated by 143 students enrolled in college-level chemistry. Findings indicated that 32% of these questions were highly evaluated by the experts.

3. Methodology

The aim of this study is to investigate how ChatGPT could be used to assess students' submitted assessments in a typography design course and how this assessment compares with the assessment provided by human evaluators. Therefore, the study adopts a mixed methods research design where the assessment reports produced by both human and AI evaluators are processed statistically and using content analysis to explore the extent to which ChatGPT can facilitate and automate the process of assessment.

3.1. Participants

3.1.1. Students

For this research, a purposive sample of 25 female students was selected. Those students were enrolled in a Typography course in the Distance Education Diploma program "Digital Media" offered at King Abdulaziz University. Profiling those students revealed that 28% of them were fresh high school graduates, and the remaining 72% spent an average of 2 years after graduating from high school. Generally, they have intermediate to advanced ICT skills, as per the requirements for success in their courses.

3.1.2. Human Evaluators

The researcher recruited two evaluators from the program "Digital Media" who teach/train students in digital media courses and who have extensive knowledge in Typography design and applications. Their job was to evaluate students' products (Typography designs) based on their knowledge of the field and the criteria provided to them by the researcher. Before the actual assessment, the assessment rubric was given to the evaluators, and it was explained to them how this rubric is in alignment with the outcomes of the course and what the students are supposed to learn.

To calibrate the performance of the evaluators, the researcher trained them on how to use the rubric's metrics to evaluate students' submissions during a two-day session. In these sessions, the researcher piloted the process of evaluation with them using three students' submissions, focusing on how to assess the three metrics (Typeface, Format, and Color) that are selected for this study.

The researcher assessed the performance of evaluators and observed a recognizable level of similarity between them.

3.2. Assessment Description and Evaluation Rubric

From an instructional design perspective, the Typography course's main learning outcomes (CLOs) are related to the exploration of visual design components considering both theoretical and practical elements (Table 1). The course requires students to produce visually appealing designs to fulfill the following learning outcomes.

Table 1. Learning Outcome Description.

Learning Outcome	Description
Develop an art piece using appropriate typographic metrics using digital means.	Primarily, this outcome focuses on enhancing learners' ability to creatively implement typographic elements in their produced artwork designs, thus motivating them to explore the aesthetic and communicative qualities of typography.
Create graphic design pieces that demonstrate the ease and familiarity with using typography in graphic design.	This outcome focuses on how the produced designs and artwork should demonstrate a level of mastery indicative of the learner's knowledge in the field.

To achieve these CLOs, a detailed assessment task was designed for students. It is important to clarify the logic behind the design of this task with reference to the domain knowledge of the program "Visual Multimedia" and the course "Typography", which is outlined below:

1. Font Selection: Students are required to include text elements (titles or descriptions) in their designs. The nuances of font selection can effectively help students understand how various typefaces can impact the tone of an artwork and influence its readability, which are considered crucial elements to achieve effective visual communication;
2. Formatting: This requirement of the assessment evaluates how students translate theoretical concepts into practical designs utilizing their critical thinking skills in choosing formatting elements;
3. Color Palette Selection: Students are expected to consider the psychological and aesthetic impact of color on their designs through an appropriate selection of color components.

The following paragraph includes the assessment requirements and elements on which students' work will be graded.

Create an engaging book cover or poster on a topic of your choice. Your design should incorporate the following elements:

Design Considerations:

- *Font Selection: Choose at least two fonts, specifying the font family, type, and size. Include images of the selected fonts from Google Fonts or any other relevant website.*
- *Formatting: Apply design principles learned in lectures, considering aspects such as harmony of elements/colours, similarity and contrasts, dominance, hierarchy, balance, and scale.*
- *Colour Palette: Select a colour palette consisting of more than two colours (gradual variations accepted).*

The next step consisted of designing a task evaluation rubric (Table 2), which serves as guidance for both students and educators in creating and evaluating the produced

artwork. The rubric includes three metrics that are used to guide students on the expected quality of the work they produce. Three important criteria are considered: the quality of typefaces, format representations, and color selection. Basically, these components are significant for effective visual communication. Each criterion is graded up to 2 points towards a total of 6. For both student and instructor, the rubric does not only quantify proficiency but also supports learners' acquisition of a comprehensive understanding of crucial design elements.

Table 2. Task evaluation rubric.

Criteria	0 Point	1 Point	2 Points
Typefaces: Evaluates the choice of typefaces, size, and readability	- Inconsistent typefaces with poor readability.	- Limited variety in typefaces, yet readable and clear. - Acceptable font choices with moderate readability.	- Excellent variety and creativity in typefaces with high readability. - High creativity demonstrated in font choices.
Format: Assesses the layout, object alignments, and overall design formatting.	- Poor layout with no clear alignment or structure. - Lacks creativity.	- Decent layout with some alignment issues. - Shows some creativity.	- Well-structured format with excellent alignment and design choices. - Shows advanced creativity.
Color: Judges the use of color in the design, including harmony and justification.	- Monochromatic or inappropriate color choice with no justification. - Poor use of color with no harmony or justification.	- Two to three colors are used with gradual transitions and some justification. - Good use of color with some harmony and justification.	- Four or more colors used with gradual transitions and suitable justification for design decisions. - Excellent use of color with perfect harmony and well-justified choices.
Total Points Possible: 6			

The grading rubric is then turned into a grading/feedback report form, which both human and AI-enhanced evaluators can utilize while analyzing the content of the students' artwork. Table 3 is a representation of the components sample of this form.

Table 3. Grading/feedback report form.

Image ID	Typeface (2 Points)	Format (2 Points)	Color (2 Points)	Total	Feedback
A	1	1	1	3	Limited typeface with some alignment issues. Lacks creativity. Good use of color with some harmony.
B	1	2	1	4	Some words are small in size with unappropriated color, which affected the readability.

3.3. Designing a Custom GPT for Automatic Assessment Evaluation

One of the main evaluation tools developed for this study is a custom ChatGPT that can assume the role of an educator and assess students' work. Basically, the research capitalizes on the capabilities of OpenAI's ChatGPT, which is a complex Large Language Model (LLM) released in November 2022. This model incorporates AI techniques to

recognize and generate text patterns that resemble ones produced by humans in precision, creativity, and variation of style. Recently, GPT-4, with its latest release in 2023, has been enhanced with internet browsing tools, image generation, and data analysis, creating, in the process, a very versatile tool.

In late November 2023, the custom GPT feature was rolled out for Plus users of OpenAI services. A user can train a custom GPT on their data, incorporating the basic capabilities of the GPT-4 like image generation, internet browsing, and data analysis to perform a specific task. Therefore, the researcher utilized the custom GPT feature to create a "Typography Evaluator" GPT, which is used to generate automatic grading and feedback on the typographical designs produced by the participating students. To achieve this objective using this custom-designed GPT, the researcher input grading the rubric in addition to an evaluation/feedback form, which must be used by the "Typography Evaluator" to grade students work considering the elements of typefaces, format, and color use, then generate suitable and informative feedback.

The design process of this custom GPT involved the following steps:

1. Objective Definition: In this first stage, the researcher defines the purpose of the "Typography Evaluator", considering that it is going to be used to assess image-based students' designs like book covers, business cards, and posters. Accessing the GPT builder, the researcher activated the "create" interface, where we added the name of the GPT "Typography Evaluator", a short description, and a set of instructions in text format (Figure 1);

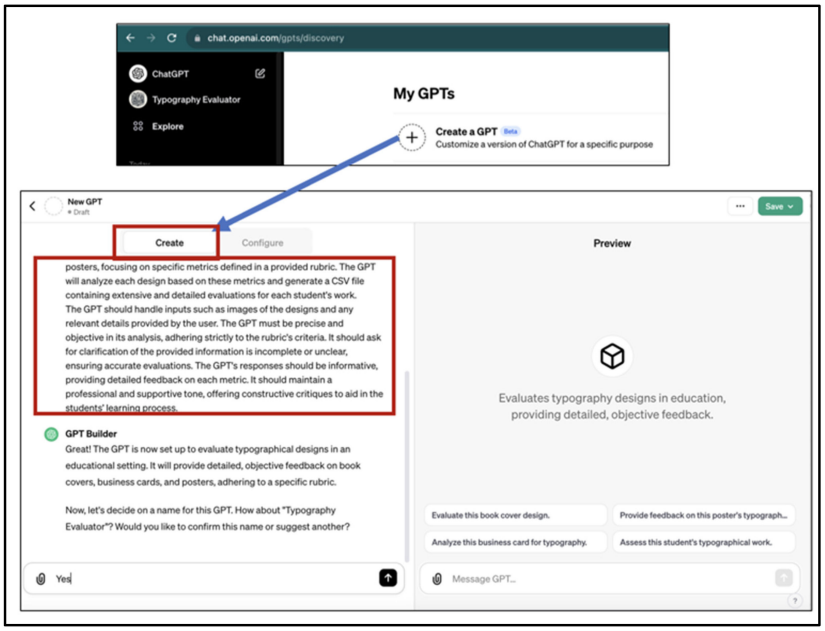

Figure 1. GPT Purpose Definition.

2. Setting up the Conversation Starters: The previously input instructors are converted automatically into conversation starters, which will help the educator interact with the GPT. Basically, these guide the educator to start the process of automatic grading and feedback based on the previously set criteria and standards used in the typography course. This will be further controlled by the extra knowledge that is going to be added to the custom GPT. Figure 2 presents some of these conversation starters;

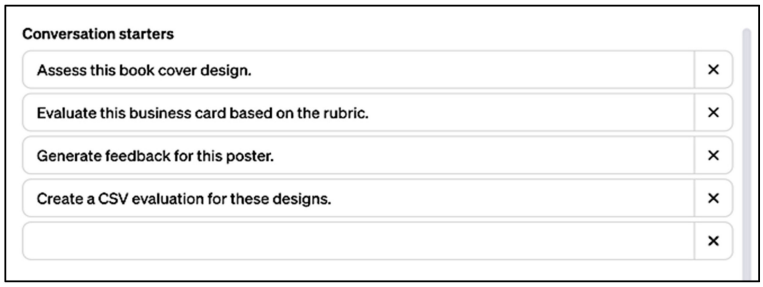

Figure 2. GPT Conversation Starters.

3. Adding Extra Knowledge: Among the most important features in the custom GPT design process is the "extra knowledge" feature, which allows users to add material in various formats (word documents, PDFs, images, etc.) to the custom GPT so that it uses this material while configuring the GPT. Both the grading rubric file and the evaluation metrics, which will be used to generate the grading/feedback report, are added to the configuration of the "Typography Evaluator" (point 3 in Figure 3);
4. Selection Required Capabilities: Customizing the GPT comes with another important feature that allows the developer of the GPT to select which tools to provide the users with. In this case, the Typography Evaluator includes capabilities such as image processing and generation using DALL-E. Two relevant capabilities were activated as well: browsing the internet and using the code interpreter, which helps in data processing and analysis. In particular, the latter capability is important for "Typography Evaluator" as it allows the computation of grading and provision of informative feedback based on data analysis (point 4 in Figure 3);

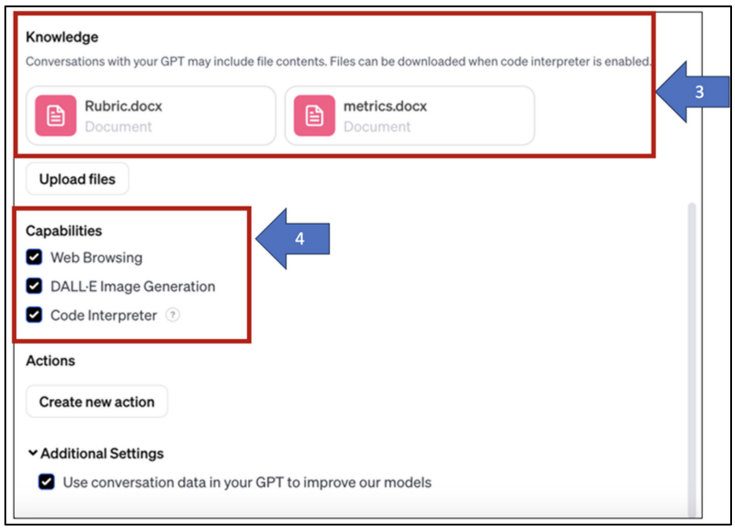

Figure 3. The third and fourth steps in creating custom ChatGPT: Arrow 3 points to the step of adding extra knowledge and Arrow 4 highlights the selection of capabilities.

5. Controls for Data Privacy: The custom GP "Typography Evaluator" is set to be accessible only via a private link, which can only be shared with specific people. In this way, the process of uploading students' artwork, grading it, commenting on it, and exporting relevant structured reports is not public, ultimately endangering students' privacy;
6. Testing the custom GPT: The "Typography Evaluator", in its final build (Figure 4), was trained on the extra knowledge data we provided and then tested for the generation of comprehensive grading and feedback of students' designs. This entails a simple deployment of the GPT and then initiating a conversation with it that includes simple and direct prompts. We started with testing the reliability of the "Typography Evalua-

tor" by uploading an image that does not classify as one of the artwork types defined for students' assessment. Also, the image does not include pre-defined elements of design in the assessment task. When promoting the "Typography Evaluator" to assess it, its response revealed that it could detect its misalignment with the assessment. The generated response: "I'm sorry, but the image you've provided is not a typographical design, and my evaluation capabilities are specific to typography, which includes aspects such as typeface, format, and colour as they relate to text" is documented in Figure 5. To further test the "Typography Evaluator" considering its feedback, the researcher used two typography designs (Figures 6 and 7) to investigate the quality of feedback and whether it is capable of distinguishing aspects of typography designs. Two book cover designs were selected. When prompting the "Typography Evaluator" to rate Figure 6, it listed among its weaknesses the following: "The staggered, colorful typography might affect readability, especially from a distance or in smaller sizes. It's also possible that the design might be too busy or vibrant for some readers, which could be a turn-off for those seeking more traditional design content". For Figure 7, the custom GPT states that the design is "The clear and direct, with strong contrast, highest readability, and a clear hierarchy of text that makes it easy to understand the book's purpose at a glance".

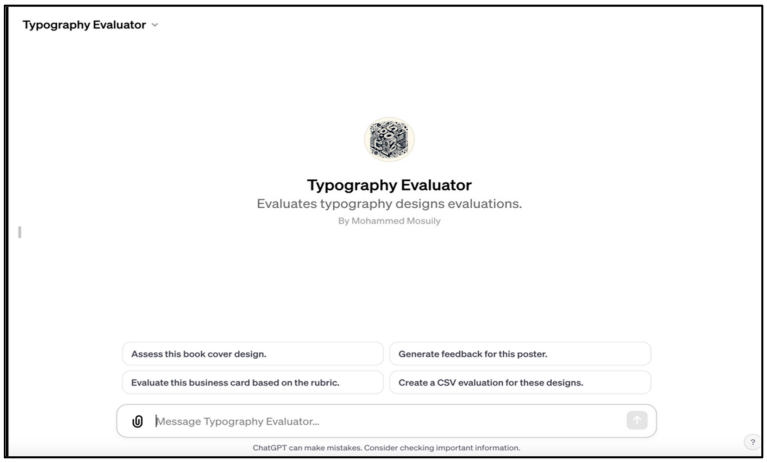

Figure 4. User view—final building step.

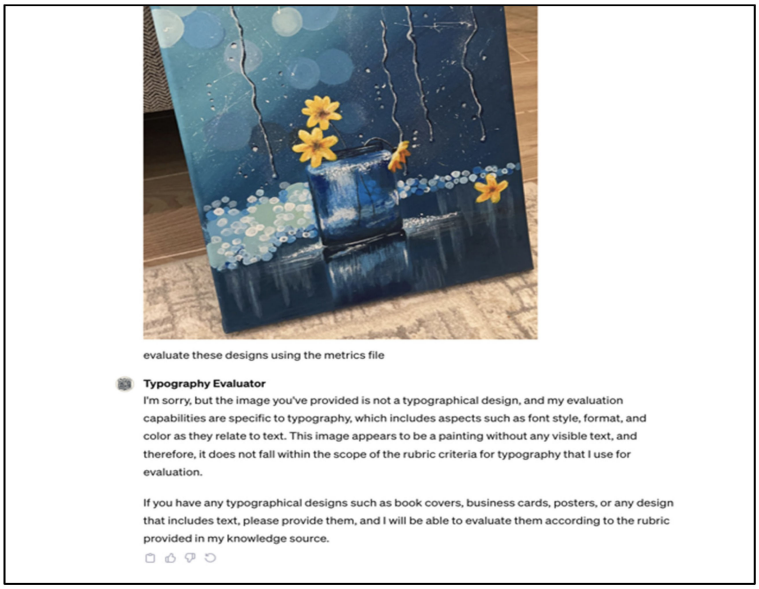

Figure 5. GPT evaluation capabilities.

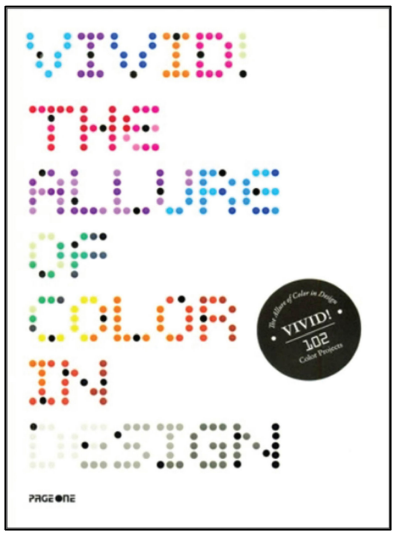

Figure 6. Ineffective book cover design.

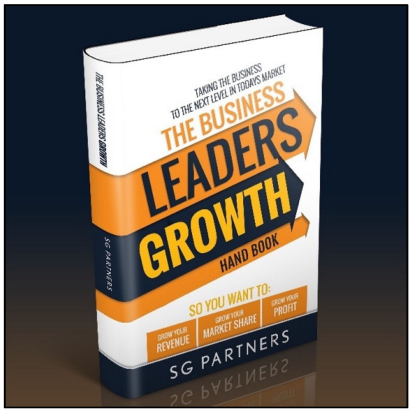

Figure 7. Effective book cover design.

7. Next, batches of 10 designs were uploaded at a time as the custom GPT possesses the capability to process only up to 10 files simultaneously. Figure 8 illustrates the three steps we follow to prompt the GPT and obtain its report in a .csv file (Figure 9).

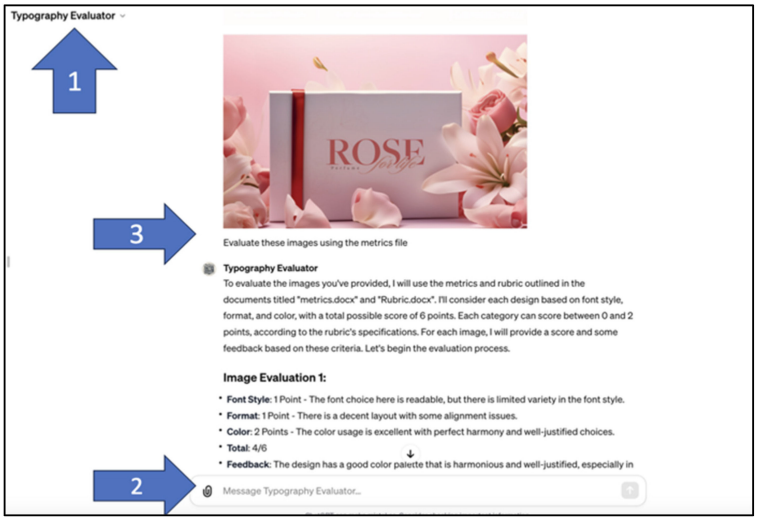

Figure 8. Guides of using Typography Evaluator interface. Select model by choosing the custom ChatGPT model (Arrow 1), Upload image for evaluation (Arrow 2) and View results in the designated area for results (Arrow 3).

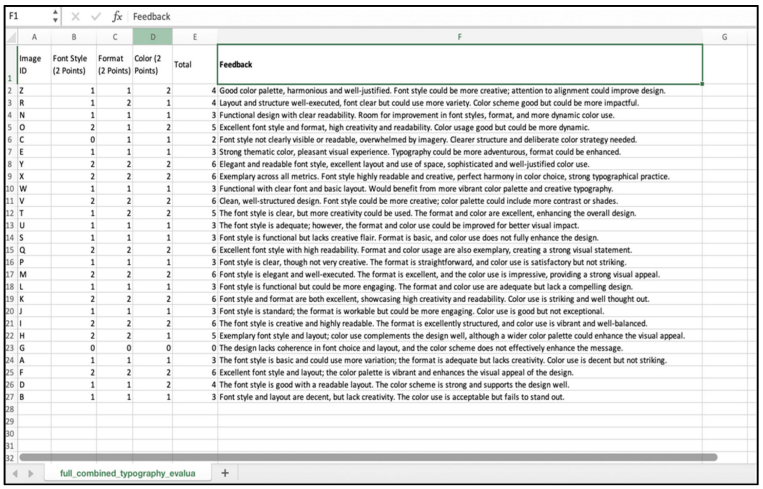

Image ID	Font Style (2 Points)	Format (2 Points)	Color (2 Points)	Total	Feedback
Z	1	1	2	4	Good color palette, harmonious and well-justified. Font style could be more creative; attention to alignment could improve design.
R	1	2	1	4	Layout and structure well-executed, font clear but could use more variety. Color scheme good but could be more impactful.
N	1	1	1	3	Functional design with clear readability. Room for improvement in font styles, format, and more dynamic color use.
O	2	1	2	5	Excellent font style and format, high creativity and readability. Color usage good but could be more dynamic.
C	0	1	1	2	Font style not clearly visible or readable, overwhelmed by imagery. Clearer structure and deliberate color strategy needed.
E	1	1	1	3	Strong thematic color, pleasant visual experience. Typography could be more adventurous, format could be enhanced.
Y	2	2	2	6	Elegant and readable font style, excellent layout and use of space, sophisticated and well-justified color use.
X	2	2	2	6	Exemplary across all metrics. Font style highly readable and creative, perfect harmony in color choice, strong typographical practice.
W	1	1	1	3	Functional with clear font and basic layout. Would benefit from more vibrant color palette and creative typography.
V	2	2	2	6	Clean, well-structured design. Font style could be more creative; color palette could include more contrast or shades.
T	1	2	2	5	The font style is clear, but more creativity could be used. The format and color are excellent, enhancing the overall design.
U	1	1	1	3	The font style is adequate; however, the format and color use could be improved for better visual impact.
S	1	1	1	3	Font style is functional but lacks creative flair. Format is basic, and color use does not fully enhance the design.
Q	2	2	2	6	Excellent font style with high readability. Format and color usage are also exemplary, creating a strong visual statement.
P	1	1	1	3	Font style is clear, though not very creative. The format is straightforward, and color use is satisfactory but not striking.
M	2	2	2	6	Font style is elegant and well-executed. The format is excellent, and the color use is impressive, providing a strong visual appeal.
L	1	1	1	3	Font style is functional but could be more engaging. The format and color use are adequate but lack a compelling design.
K	2	2	2	6	Font style and format are both excellent, showcasing high creativity and readability. Color use is striking and well thought out.
J	1	1	1	3	Font style is standard; the format is workable but could be more engaging. Color use is good but not exceptional.
I	2	2	2	6	The font style is creative and highly readable. The format is excellently structured, and color use is vibrant and well-balanced.
H	2	2	1	5	Exemplary font style and layout; color use complements the design well, although a wider color palette could enhance the visual appeal.
G	0	0	0	0	The design lacks coherence in font choice and layout, and the color scheme does not effectively enhance the message.
A	1	1	1	3	The font style is basic and could use more variation; the format is adequate but lacks creativity. Color use is decent but not striking.
F	2	2	2	6	Excellent font style and layout; the color palette is vibrant and enhances the visual appeal of the design.
D	1	1	2	4	The font style is good with a readable layout. The color scheme is strong and supports the design well.
B	1	1	1	3	Font style and layout are decent, but lack creativity. The color use is acceptable but fails to stand out.

Figure 9. Typography Evaluator Output.

3.4. Data Collection and Analysis

3.4.1. Data Collection

The submitted students' artwork is sent to two human evaluators who are experts in the field of design, alongside the grading rubric and the grade/feedback report form.

Both evaluators were required to provide their candid assessment of students' designs based on the rubric criteria considering the division of grades. A two-day assessment cycle was initiated for both evaluators to process and analyze the content and grade 25 artworks. An average of 10 min per design was spent assessing the artwork and producing the grade/feedback for each student. This is approximately a 4-h task for each evaluator. By the end of this cycle, they submitted their reports for each student.

The automatic grade/feedback cycle with the "Typography Evaluator" lasted for 30 min, spent on uploading the 25 images (10 per batch) and then generating and exporting the grade/feedback report in .csv format. This evaluation process resulted in three assessment reports, including the grading for each design criterion, a total grade, and a feedback statement.

3.4.2. Analysis Techniques

To investigate the research questions, we will conduct two statistical tests on the evaluation data coming from the GPT-based evaluator and two human evaluators. The Shapiro–Wilk test will be implemented to check the normality of distribution for each dataset based on the scores from each evaluator. The assessment of normality of data distribution is important because it will impact the selection of the statistical tests. Basically, parametric tests, such as the *t*-test, demand normally distributed data values, whereas non-parametric tests do not. The generated Shapiro-Wilk should be above the standard level of 0.05 to indicate a normal distribution. Subsequently, the selection of either a *t*-test or a Mann–Whitney U test will be decided. In addition, the inter-rater reliability is also computed using Cohen's Kappa statistic to represent the level of consistency between the two evaluators. Cohen's Kappa values range from −1 (indicative of the absence of inter-rater reliability) to 1 (indicative of perfect inter-rater reliability), which is a standardized method to evaluate the extent to which the human evaluators' assessments align beyond consideration of chance.

To investigate the educators' as well as automatically generated textual feedback, the researcher conducted quantitative content analysis to identify the most frequent themes or topics in these feedback snippets [19]. Basically, the first step in this structured approach entails creating several overall categories (or themes) that will be the focus of the investigation. Then, the numerical frequencies for those themes as word representation are computed using ChatGPT for both the human evaluators and the Typography Evaluator.

Finally, a comparison of each category representation per evaluator is conducted to discover similarities and differences in the quality of offered feedback.

4. Results

To address the research questions, various analyses were conducted on both the quantitative (grades) and qualitative (feedback) data that each of the evaluators (human or AI-based) had produced during the process of evaluating students' artworks.

In this section, these results are presented and explored, considering their relation to the research questions previously defined. First, the results based on the statistical investigation are introduced, and then the ones that quantify the qualitative data are discussed as well.

4.1. Testing Normality of Distribution

As the scored elements in the evaluation reports are numerical values contributing to a total grade, the initial stage of the analytical process adopted in this research highlighted the importance of testing the normality of data distribution for each evaluator. This step determines the type of statistical techniques to be used to test the significance of differences between evaluators (Figure 10). So, the Shapiro–Wilk test is implemented on each evaluator's dataset to compute the *p*-value. Ideally, if this value is below the alpha threshold of 0.05, this would indicate a significant divergence from normality. When applied in the research context, it revealed that the total scores coming from the reports of all evaluators did not follow a normal distribution. Table 4 shows the *p*-values for Evaluator 1, Evaluator 2, and Typography Evaluator, which scored 0.0008, 0.049, and 0.0027, respectively. This demands the implementation of non-parametric tests in further analyses.

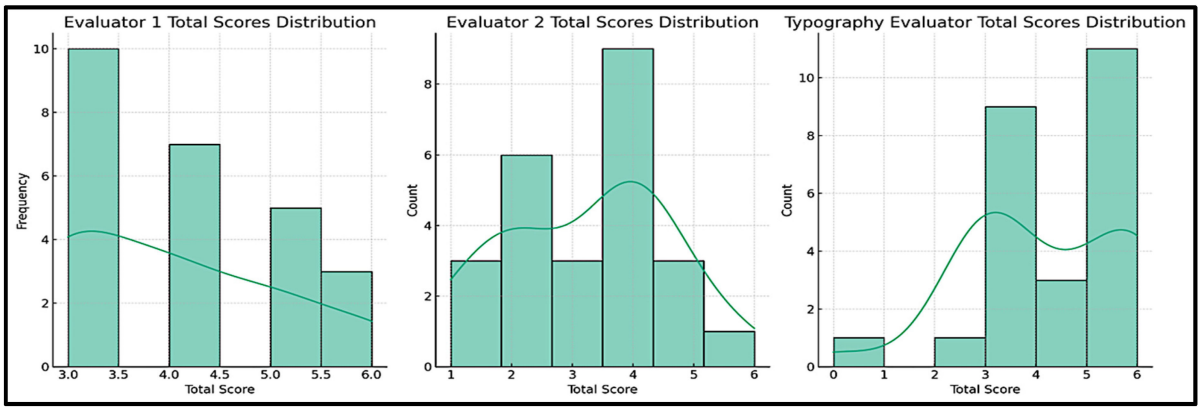

Figure 10. A histogram plot is utilized to compare data distribution among evaluators and the Typography Evaluator, assessing the assumption of normality.

Table 4. Results of Shapiro–Wilk test.

Evaluator	Shapiro–Wilk Statistic	*p*-Value	Normal Distribution
Evaluator 1	0.8317	0.0008	No
Evaluator 2	0.9193	0.0493	No
Typography Evaluator	0.8598	0.0027	No

4.2. Inter-Rater Reliability

To investigate inter-rater reliability between the results provided by both human evaluators, the researcher computed the Cohen's Kappa statistic based on the rubric's total out of 6. Cohen's Kappa value (0.4024) is within the range of 0.4 to 0.6, which is classified as a moderate form of reliability indicative of a significant level of consensus in their assessments of students' work.

4.3. Differences in Grading Patterns between Evaluators Based on Total Scores

To address the first question in the current study: "Is there a significant difference in the grading patterns between AI-based and human evaluations of students' projects?" the researcher utilized the Mann–Whitney U test to investigate the significance of mean differences between the three evaluations based on the total scores achieved by each student in the evaluation process. The Mann–Whitney U test is useful in this case because it allows for the comparison of two independent samples where assumptions of normal distribution are not met. Using this test, the ranks of the scores are computed rather than their absolute values.

The outcomes of the test, as shown in Table 5 and Figure 11, denote that the *p*-value for the comparison between Evaluator 1 and the Typography Evaluator is approximately 0.701, which is indicative of the absence of significant variance in the evaluators' grading patterns. However, the mean comparison of Evaluator 2 and Typography Evaluator yielded a ≈ 0.047 *p*-value, highlighting the presence of a statistically significant difference in the scoring mechanisms of these two evaluators.

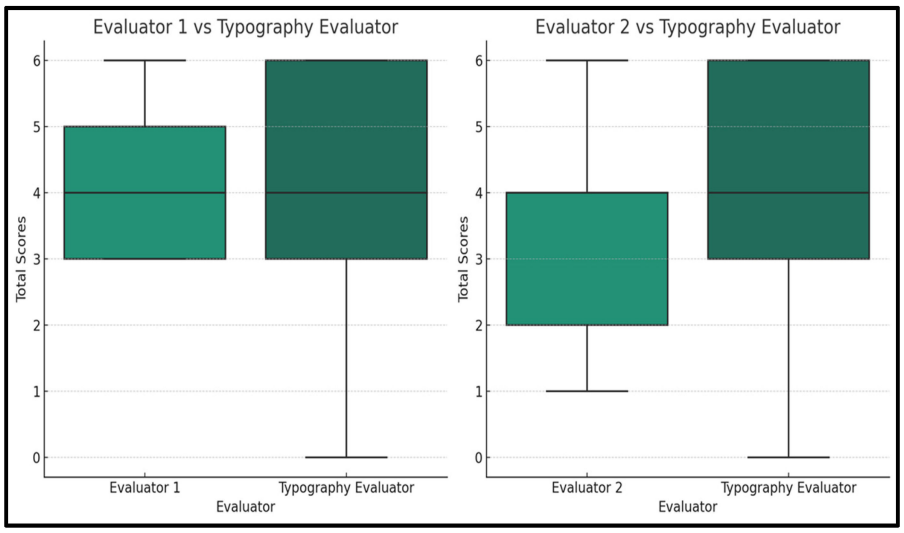

Figure 11. Box plot shows Mann-Whitney U test score for Evaluator Comparison based on Total Scores.

Table 5. Mann–Whitney U-test (Grading Patterns between evaluators and Total scores).

Comparison	Mann–Whitney U Statistic	*p*-Value	Significant Difference
Evaluator 1 vs. Typography Evaluator	293	0.701143	No
Evaluator 2 vs. Typography Evaluator	211.5	0.046822	Yes

4.4. Differences in Grading Patterns between Evaluators Based on Rubric Criteria Scores

To further investigate the significance of variance in the evaluators' grading based on the total score for each assessed design aspect: typeface, format, and color, several Mann–Whitney U tests are conducted. The aim has been the comparison of each human evaluator's scores with those of the Typography Evaluator. This is performed to address the second research question: "Is there a significant difference in the grading patterns for each design aspect between the AI-based and human evaluations of students' projects?"

The results of the tests revealed that for the "typeface" aspect, there is no statistical significance between the three evaluations because the comparisons' *p*-values between Evaluator 1 and the Typography Evaluator (=0.865) and Evaluator 2 versus the Typography Evaluator (=0.309) are above the 0.05 threshold.

The test results for the "format" aspect indicated no significant differences between Evaluator 1 and the Typography Evaluator ($p = 0.681$). However, for Evaluator 2's comparison to the Typography Evaluator, the *p*-value of 0.007 indicates a significant differ-

ence between them in terms of their assessment criteria when evaluating this specific design aspect.

The absence of statistical significance is recorded as well when investigating the aspect of "Colour". The computed *p*-values were 0.136 for Evaluator 1 and 0.113 for Evaluator 2 when compared with those of the Typography Evaluator, indicating that there are no significant differences in their assessment criteria.

Considering the insights reported in Table 6 and Figure 12, it could be assumed that both human and AI-based evaluations showcase mostly similar scoring mechanisms for the design aspects of typeface and color in the assessed projects. It is only when considering the element of format assessment by Evaluator 2 in comparison to the Typography Evaluator that one notes a significant difference, which is indicative of evaluation variability.

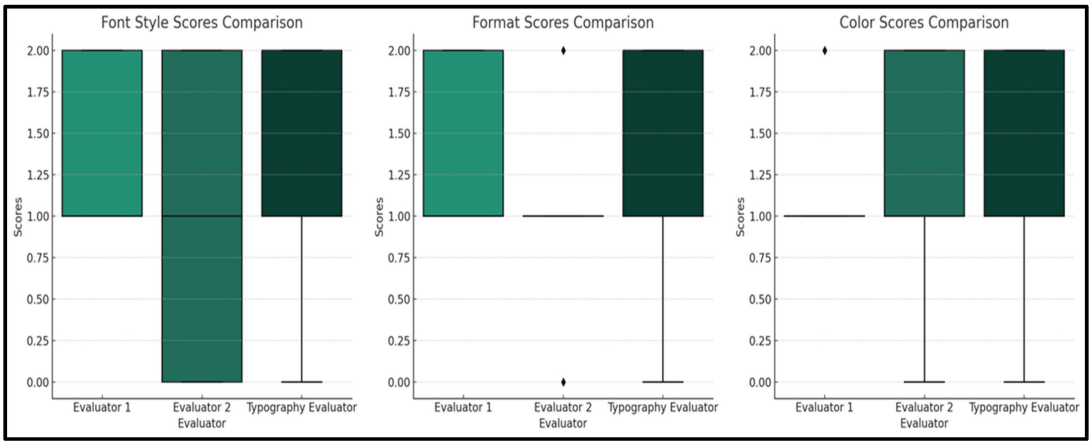

Figure 12. Comparative Analysis of Grading Patterns, based on rubric criteria scores Typeface, Color and Format Assessment Across Human and AI Evaluators.

Table 6. Mann–Whitney U test (Grading Patterns between evaluators based on rubric criteria scores).

Design Aspect	Comparison	Mann–Whitney U Statistic	*p*-Value	Significant Difference
Typeface (Font style)	Evaluator 1 vs. Typography Evaluator	304.5	0.8654	No
	Evaluator 2 vs. Typography Evaluator	263.5	0.3087	No
Format	Evaluator 1 vs. Typography Evaluator	331.5	0.681	No
	Evaluator 2 vs. Typography Evaluator	197.5	0.0066	Yes
Color	Evaluator 1 vs. Typography Evaluator	247	0.1359	No
	Evaluator 2 vs. Typography Evaluator	239.5	0.1129	No

4.5. Differences in Feedback Provided by Evaluators

To answer the research's third question, "Does the written feedback provided by the human evaluators significantly differ from that of AI-based evaluator in terms of comprehensiveness across various design criteria in students' projects?" the researcher conducted a significance test that utilizes the Chi-Square test of independence to compare the frequencies of specific feedback words aligned to a set of pre-defined categories. This test is suitable for comparing frequencies across different groups and can help determine if there are statistically significant differences in the distribution of feedback among the evaluators.

The process starts with a basic content analysis identification of six categories and using ChatGPT to compute the frequencies of words that relate to these categories for each evaluator Figure 13 and Table 7.

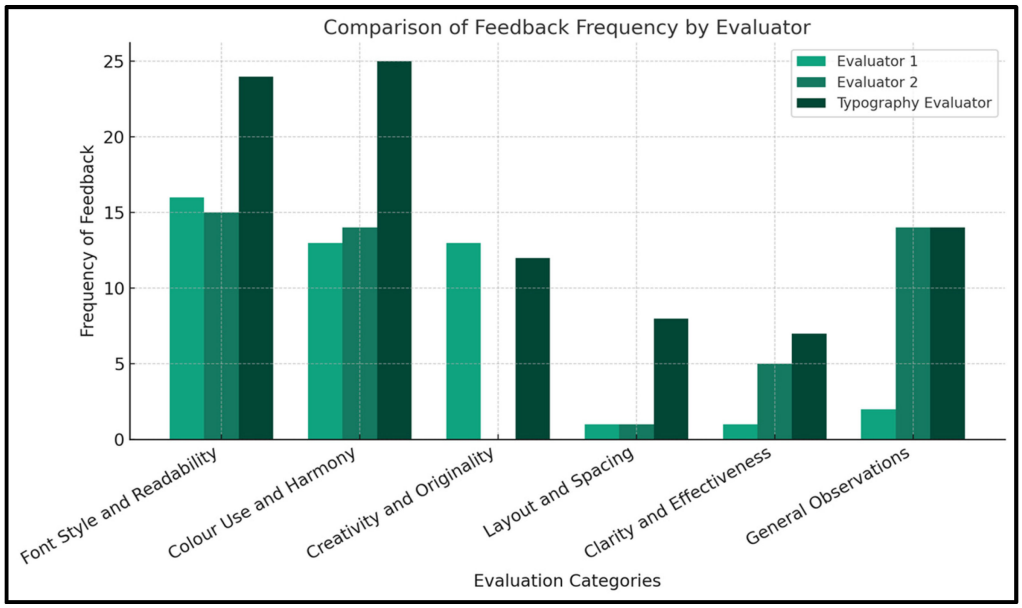

Figure 13. All evaluator feedback comparison.

Table 7. Evaluator feedback comparison.

Category	Frequency Evaluator 1	Frequency Evaluator 2	Frequency Typography Evaluator
Font style and Readability	16	15	24
Colour Use and Harmony	13	14	25
Creativity and Originality	13	0	12
Layout and Spacing	1	1	8
Clarity and Effectiveness	1	5	7
General Observations	2	14	14

Then, the Chi-Square test of independence and the related *p*-value are computed based on the category frequency per evaluator data. The computed Chi-Square statistic is approximately 29.61 with a *p*-value of 0.00099, which indicates statistically significant differences in the way the feedback is distributed on the six pre-defined categories per evaluator. This could be read to denote that the comprehensiveness of feedback words provided by the human evaluators differs from that of the Typography Evaluator.

5. Discussion

The quantified results presented previously highlight how effective an AI solution can be in the context of academic assessment and how important it is to consider areas of improvement and potential. For example, when considering the differences in grading based on the total scores of students, one noted the absence of significant differences in grading mechanisms between Evaluator 1 and the "Typography Evaluator", but a significant difference when compared to Evaluator 2. The first case suggests that AI is both consistent and objective, with a performance that is comparable even to humans. This underscores the effectiveness of the custom GPT AI and the potential of using it, at least for preliminary assessments, allowing more time for the educator to deal with the more nuanced or complex elements of the student's design. The diverse opinion offered by the Typography Evaluator in comparison to the second evaluator might indicate a return to a more subjective assessment perspective, but one might argue as well that it might reflect the importance of standardizing the assessment task criteria in ways that contribute to more successful alignment between the AI and human evaluators. These differences,

one can assume as well, reflect not only on the subjective elements of artistic evaluation that AI is still learning to grasp but also point out the importance of assessment criteria standardization and calibration.

When considering how the Typography Evaluator performed against both human evaluators regarding its performance on rubric categories: font, format, and color, one can safely point out that the AI solution's ability to align with both evaluators while assessing the font and color categories is remarkable. Such observed alignment denotes that both human and AI evaluators can demonstrate comparable levels of assessment of fine-grained standardized aspects of students' designs, pinpointing the Typography Evaluator's potential to reliably support human judgment in pre-defined areas.

The significant differences when it comes to evaluating students' artwork based on format elements showcase the potential for refining the rubric criteria by incorporating elements like context, artistic intent, and aesthetic impact so that all evaluators can calibrate their assessment accordingly. The differences in the significance of the frequency distribution of the text-based feedback provided by evaluators offer as well significant insights into the volume and breadth of such feedback AI, hence its comprehensiveness. The AI evaluator seems to provide more diverse words of feedback that are related to the categories when compared to human evaluators.

When it comes to human feedback, one noted that in some instances, feedback is not provided, and at other times, it is not detailed enough to focus on all the elements provided in the rubric. On the other hand, the Typography Evaluator provides text feedback that is consistent, objective, and closely related to specific criteria wording that has been added to its system as extra knowledge. Moreover, the human feedback included words that were more of an overall judgment on the design as ineffective or lacking, whereas the feedback extracted from the AI focused on using words free of bias and was uniformly implemented throughout all students.

A qualitative reading of the various feedback comments provided by the three evaluators is similarly important as it would clarify the points of alignment between the AI and human evaluators. Both human evaluators have demonstrated agreement on the type of terms used in the feedback. For example, in harmonious color use, both evaluators agreed on utilizing terms related to color choices, creativity, clarity, and effectiveness. Such words as "alignment", "bad", "choosing", "creativity", and "good" denote this attention to creativity details. Considering layout and spacing, both evaluators assessed visual aspects with very limited feedback using terms like "colour", "font", and "style".

The Typography Evaluator, on the other hand, used a set of terms denoting a distinct and broader perspective in assessing the same design aspects. Considering color use, the Typography Evaluator presents a richer terminology in its feedback, like "adventurous", "benefit", and "colours" in contexts that are appreciative of creative and diverse colour schemes. Feedback on the layout and spacing aspects is more comprehensive, pinpointing overall visual appeal, clarity, and spatial elements, which were not highly emphasized by the human evaluators. Generally, there is an alignment in the use of terms and feedback descriptions between the AI and Evaluator 1, especially when considering font and style assessment.

6. Limitations

Generally, the findings of the current research are informative and support the research's initial argument that custom GPTs can be trained as assessment tools. Despite these findings, we have to observe the restricted scope of the experiment reported in this study, in which the feedback of the "Typography Evaluator" has been exclusively compared with the feedback of only two human evaluators. We acknowledge that the dependence on a limited number of evaluators may introduce subjectivity and biases into the assessment process; consequently, the results of this study should be considered within the context of an initial investigation of the capabilities of ChatGPT in educational assessment in the form of a case study that might be expanded on in the future.

7. Conclusions

Investigating the potential contributions of generative AI and LLMs like ChatGPT and GPT-4 and the continuous stream of new features and plugins is a necessity in the field of educational assessment. We tried to present an initial examination of such potential by presenting an AI-enhanced evaluation tool, "Typography Evaluator", which utilizes OpenAI's GPT-4 and the custom GPT feature. The researcher used this tool to automate the process of evaluation of students' image-based typography designs. Some of the findings indicate the presence of significant differences between the tool's performance and human evaluators', but mostly, similar practices and outcomes are reported.

Further research in this area might consider allowing students access to the custom GPT so that they can use it for self-assessment prior to submitting their work, as well as evaluate the feedback offered to them via the solution. Moreover, researchers might consider the provision of multilingual or multimodal feedback mechanisms conducted using such tools as the Typography Evaluator. The recruitment of more human evaluators is a very important direction in this area of research, as well as the utilization of human coders who will identify the most prominent topics or themes in both the human and AI evaluators' textual feedback.

In addition, a collaborative assessment system might incorporate this Typography Evaluator tool in conjunction with an interface where human evaluators participate in the evaluation process. It is assumed that such a system will increase fairness and decrease subjectivity and bias in assessment. The differences and similarities in grading patterns observed between Evaluator 2 and the Typography Evaluator offer a multifaceted perspective on the integration of AI in educational assessment, particularly in artistic fields.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the reported results of this study are available on: https://github.com/Malmasre/Typography-results.git (accessed on 26 November 2023).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Cockburn, I.M.; Henderson, R.; Stern, S. The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis. In *The Economics of Artificial Intelligence: An Agenda*; University of Chicago Press: Chicago, IL, USA, 2018; pp. 115–146. Available online: https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/impact-artificial-intelligence-innovation-exploratory-analysis (accessed on 26 November 2023).
2. Gillani, N.; Eynon, R.; Chiabaut, C.; Finkel, K. Unpacking the 'Black Box' of AI in Education. *Educ. Technol. Soc.* **2023**, *26*, 99–111.
3. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]
4. Amri, M.M.; Hisan, U.K. Incorporating AI Tools into Medical Education: Harnessing the Benefits of ChatGPT and Dall-E. *J. Nov. Eng. Sci. Technol.* **2023**, *2*, 34–39. [CrossRef]
5. Muñoz, A.M. Creating Persuasive Infographics for Advocacy Groups in an e-SL Language Course. In Proceedings of the 2022 De La Salle University Research Congress, Manila, Philippines, 6–8 July 2022.
6. Fu, Z.; Wang, Y. Analysis of Aesthetic Principles in Typography. *Trans. Soc. Sci. Educ. Humanit. Res.* **2023**, *1*, 182–186.
7. Shen, J.J.; Jin, K.; Zhang, A.; Breazeal, C.; Park, H.W. Affective Typography: The Effect of AI-Driven Font Design on Empathetic Story Reading. In *Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, in CHI EA'23, Hamburg, Germany, 23–28 April 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–7. [CrossRef]
8. Wiggins, G.P. *Assessing Student Performance: Exploring the Purpose and Limits of Testing*; Jossey-Bass/Wiley: Hoboken, NJ, USA, 1993; p. 316.
9. MacGinitie, W.H. Some Limits of Assessment. *J. Read.* **1993**, *36*, 556–560.
10. Thorp, H.H. ChatGPT is fun, but not an author. *Science* **2023**, *379*, 313. [CrossRef] [PubMed]
11. Mao, R.; Chen, G.; Zhang, X.; Guerin, F.; Cambria, E. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. *arXiv* **2023**, arXiv:2308.12488. [CrossRef]
12. Frieder, S.; Pinchetti, L.; Griffiths, R.-R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.C.; Chevalier, A.; Berner, J. Mathematical Capabilities of ChatGPT. *arXiv* **2023**, arXiv:2301.13867. [CrossRef]

13. Haleem, A.; Javaid, M.; Singh, R.P. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Trans. Benchmarks Stand. Eval.* **2022**, *2*, 100089. [CrossRef]
14. Koubaa, A.; Boulila, W.; Ghouti, L.; Alzahem, A.; Latif, S. Exploring ChatGPT Capabilities and Limitations: A Survey. *IEEE Access* **2023**, *11*, 118698–118721. [CrossRef]
15. Su, J.; Yang, W. Unlocking the Power of ChatGPT: A Framework for Applying Generative AI in Education. *ECNU Rev. Educ.* **2023**, *6*, 355–366. [CrossRef]
16. Cope, B.; Kalantzis, M.; Searsmith, D. Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educ. Philos. Theory* **2021**, *53*, 1229–1245. [CrossRef]
17. Fijačko, N.; Prosen, G.; Abella, B.S.; Metličar, Š.; Štiglic, G. Can novel multimodal chatbots such as Bing Chat Enterprise, ChatGPT-4 Pro, and Google Bard correctly interpret electrocardiogram images? *Resuscitation* **2023**, *193*, 110009. [CrossRef]
18. Memarian, B.; Doleck, T. ChatGPT in education: Methods, potentials, and limitations. *Comput. Hum. Behav. Artif. Hum.* **2023**, *1*, 100022. [CrossRef]
19. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
20. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [CrossRef]
21. Naismith, B.; Mulcaire, P.; Burstein, J. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Toronto, ON, Canada, 13 July 2023*; Kochmar, E., Burstein, J., Horbach, A., Laarmann-Quante, R., Madnani, N., Tack, A., Yaneva, V., Yuan, Z., Zesch, T., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 394–403. [CrossRef]
22. Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; Chen, G. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, 10–13 July 2023. [CrossRef]
23. Moore, S.; Nguyen, H.A.; Bier, N.; Domadia, T.; Stamper, J. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*; Hilliger, I., Muñoz-Merino, P.J., De Laet, T., Ortega-Arranz, A., Farrell, T., Eds.; Lecture Notes in Computer Science Book Series; Springer International Publishing: Cham, Switzerland, 2022; pp. 243–257. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.