Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 05 May-2026, Page No.-10069-10083

DOI: 10.47191/etj/v11i05.42, I.F. – 8.595

Monroe Grading Agent: AI-Powered Automated Assignment Grading in Higher Education

Sahar Bukhari[1], Ebenezer Amakeh[2]

[1,2]KGS Academic Center, Monroe University

ABSTRACT: Automated grading has emerged as a valuable approach to address the heavy workload and consistency challenges of manual grading in higher education. Instructors often spend countless hours evaluating student assignments, which can delay feedback and introduce subjective inconsistencies. Prior research highlights the efficiency and objectivity of AI-assisted grading, noting its potential to reduce grading inconsistencies by up to 44% (Yigit et al., 2024) and deliver faster, more scalable feedback (Wang et al., 2022). This paper introduces the Monroe Grading Agent (MGA) – an AI-powered desktop application developed to automatically grade student semester assignments by comparing submissions against the assignment instructions. MGA utilizes TF-IDF vectorization, cosine similarity, and a novel cube-root scaling method to provide fairer, partial-credit-based grading. Evaluated using realistic Monroe University assignment data, MGA demonstrates significant improvements in grading efficiency, scoring consistency, and feedback richness. A before-and-after analysis shows that MGA can drastically reduce grading time while maintaining scores comparable to human evaluators. The system also generates structured feedback aligned with assignment criteria. The paper features GUI screenshots, contextualizes MGA within the broader automated grading literature, and explores implications for teaching faculty. It concludes by outlining future enhancements including LMS integration, rubric-based scoring, and support for programming assignments, positioning MGA as a practical and scalable solution for modern academic environments.

1. INTRODUCTION

Grading students' work is widely recognized as a **time-consuming** and often challenging task for faculty, as it demands not only meticulous evaluation of each submission but also a high degree of consistency and fairness. Instructors must carefully interpret varying levels of student understanding, **ensure objectivity in assessment**, and provide timely, **constructive feedback**—all while managing large class sizes and tight academic schedules. The cognitive load and emotional fatigue associated with prolonged grading periods can impact the quality and impartiality of assessments. These challenges highlight the growing need for innovative solutions, such as **AI-powered grading systems**, to support educators in maintaining high standards of evaluation while improving efficiency and scalability.

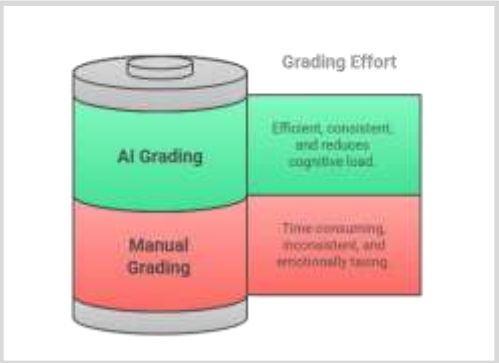

Figure 1.1. Grading Effort: From High burden to AI-Assisted Efficiency.

1.1. The Burden of Manual Grading in Higher Education

In higher education, instructors can spend dozens of hours marking assignments, which not only strains instructor workload but also delay feedback to students. For example, a new faculty member reported spending over 80 hours to grade 80 three-page papers – essentially one hour per paper. Such extreme cases underscore how traditional manual grading burdens educators' time and may limit the frequency and depth of assignments given. Research has noted that **Automated Essay Scoring (AES)** and evaluation tools can alleviate this burden by reducing the time teachers spend on grading and providing feedback. By speeding up

the grading process, instructors could assign more practice opportunities without overwhelming themselves. Moreover, students benefit from receiving more immediate – and more consistent – feedback, which better supports iterative learning and revision.

1.2. The Emergence of AI-Powered Grading

Calls to integrate **Artificial Intelligence (AI)** in post-secondary education for tasks like grading are growing, as the technology promises to enhance efficiency and consistency. **AI-powered grading systems** can analyze student work at scale and return results quickly, giving students timely insight into their performance. For instance, AI tools can sometimes grade objective responses near-instantly, and even for essays or reports, AI can highlight relevant content and provide preliminary evaluations much faster than a human grader. This immediacy can free instructors to focus on higher-order teaching tasks such as mentoring and curriculum development. Additionally, AI grading algorithms apply the same criteria uniformly across all submissions, potentially offering more consistent scoring and unbiased evaluation compared to individual human graders. Consistency is crucial for fairness – a human grader's fatigue or subjective bias can lead to variations in scores, whereas an AI agent will evaluate the same content in the same way every time. The claim that AI use in grading can reduce grading inconsistencies by up to 44% relative to human grading is supported by a study conducted by Gobrecht et al. (2024). In their research, the authors developed an AI-based automatic short answer grading (ASAG) system and compared its performance to that of certified human graders. The study found that the AI model's median absolute error was 44% smaller than that of human re-graders when compared to official historic grades, indicating that the AI system was more consistent in grading than human evaluators

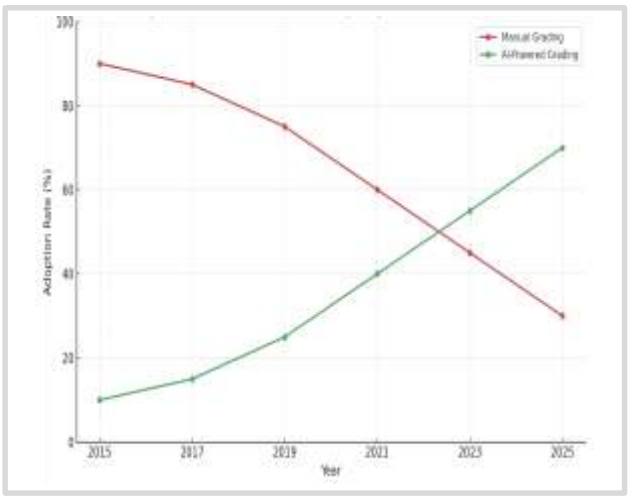

Figure 1.2. Growth of AI-Powered grading in Higher Education (2015-2025).

1.3. Challenges and Ethical Considerations in AI Grading

Despite these potential benefits, educators are appropriately cautious about fully delegating assessment to AI. Not all assignments are well-suited for AI grading – especially openended, creative projects where nuance and qualitative judgment are key. There are also concerns regarding the fairness and ethics of algorithmic grading. An AI model trained on limited or biased data could inadvertently favor certain writing styles or content, raising questions of bias and academic integrity. Furthermore, commercial AI grading services may introduce privacy and cost issues. Therefore, the consensus in the literature is that AI grading tools should *complement* rather than replace human instructors' judgment. Used responsibly, AI can handle the repetitive aspects of grading and provide consistent baseline feedback, while instructors remain involved for deeper evaluation and personal guidance.

1.4. The Monroe Grading Agent (MGA): A Practical Solution

Within this context, the *Monroe Grading Agent (MGA)* was developed as a practical AIbased grading assistant for a university setting. The goal of MGA is to streamline the grading of student assignments by automatically comparing each student's submitted work to the assignment instructions or rubric provided by the instructor. The system is built on established information retrieval techniques – namely TF-IDF (Term Frequency– Inverse Document Frequency) text vectorization and cosine similarity – to measure how well a submission's content aligns with the expected content described in the instructions. A custom cube-root scaling is applied to the raw similarity scores to produce final grades in a manner that rewards partial fulfillment of requirements more fairly. By design, the MGA aims to produce scores that are more consistent across different students and different graders, reduce the turnaround time for grading, and provide each student with feedback on how well they addressed the assignment criteria.

1.5. Research Objectives and Scope of the Paper

This paper presents the architecture of the Monroe Grading Agent and an evaluation of its impact on grading efficiency, consistency, and feedback quality. We incorporate sample data from Monroe University's student semester assignments to compare the traditional manual grading process with the automated grading process. A before-and-after analysis examines key metrics including time taken to grade, variation in scoring, and the richness of feedback given to students. Figure 1 and Figure 2 illustrate the MGA's user interface, including the grading input screen and the results output screen, respectively. In the following sections, we review related work on automated grading and AI in education, describe the design and methodology behind MGA, present comparative results with synthetic data, and discuss the implications for instructors. Possible

improvements and future directions – such as integration with Learning Management Systems (LMS), incorporation of rubric-based scoring, and support for code assignments – are also discussed, with the aim of guiding further development of AI tools that enhance teaching and learning.

2. LITERATURE REVIEW

2.1. AI in Automated Grading: Benefits and Challenges

The idea of using AI to grade student work has been explored for several decades and has gained significant traction in recent years as algorithms and computing power have improved. **Automated grading systems** range from simple programs that check answers against a key to complex machine learning models that evaluate essays or open-ended responses. The literature consistently highlights several key benefits of AI-assisted grading. First and foremost is **efficiency**: AI can dramatically reduce the time educators spend on grading, especially for large classes or repetitive assignments. By automating the scoring process, instructors can reallocate time to instructional design, one-on-one student support, or research. For example, Xiang et al. (2024) noted that manual assessment is "time-consuming, labor-intensive, and error-prone," whereas replacing it with automatic assessment reduces teacher workload and provides **instant feedback** to students.

This **timely feedback** is the second major benefit. Multiple sources emphasize that students receive feedback much faster with automated systems, which enhances learning outcomes. Wang et al. (2022) found that automated writing evaluation systems enabled teachers to assign more writing tasks and gave students immediate and consistent feedback, supporting iterative revision cycles.

Another widely cited advantage of AI grading tools is **consistency and objectivity** in scoring. Automated Essay Scoring (AES) tools apply the same criteria uniformly, avoiding inconsistencies introduced by human graders' subjectivity or fatigue. For large-scale assessments, this leads to fairer evaluations. Zou et al. (2023) observed that AI grading can closely approximate human grading while reducing score variance. Yigit et al. (2024) highlighted that AI use in grading reduced inconsistencies by up to 44%, minimizing bias across student submissions. This consistency increases **trust and fairness**, as each student is judged against the same standard. Barnett et al. (2023) similarly emphasized that AI scoring eliminates grader fatigue and subjective variation.

In addition, automated grading systems can offer **detailed feedback** at scale. Platforms like Gradescope have been used to group similar answers and apply rubric-based comments, streamlining the feedback process. Wang et al. (2022) distinguished that while simpler systems might only provide a score or basic keyword-based comments, more advanced tools can offer summary feedback on strengths and weaknesses. Zou et al. (2023) further discussed how such systems can point out grammar and structure issues, guiding students in improving their drafts.

Despite these advantages, researchers emphasize several **challenges and limitations**. One major concern is that AI grading may be **too superficial** for creative, open-ended, or higher-order thinking tasks. Algorithms based on keyword matching may miss the nuance in argumentation, originality, or contextual appropriateness. Barnett et al. (2023) warned that an AI might assign high scores to essays filled with relevant terms but lacking substance, or penalize unconventional yet insightful submissions. Kumar et al. (2023) echoed this concern, emphasizing that AI scoring works best when combined with human judgment for more subjective tasks.

Bias and fairness are also major concerns. Silvestrone and Rubman et al. (2024) warned that AI models trained on skewed datasets may inadvertently favor specific writing styles or content. For instance, essays written in non-standard dialects or from underrepresented perspectives may be unfairly scored. Zou et al. (2023) also highlighted how historical biases in training data could cause AI tools to undervalue certain topics or demographics. These findings underscore the importance of **bias mitigation, transparency**, and routine auditing of AI grading tools. Silvestrone and Rubman et al. (2024) recommended making grading algorithms more interpretable and allowing faculty oversight.

Other **practical concerns** involve privacy, data protection, and user acceptance. Barnett et al. (2023) pointed out that uploading student work to commercial AI grading platforms could raise privacy issues, especially with proprietary or sensitive content. Kumar et al. (2023) discussed resistance from both faculty and students who view machine grading as impersonal or potentially inaccurate. Silvestrone and Rubman et al. (2024) suggested clear opt-in policies, instructor override functionality, and transparent communication as best practices for increasing acceptance.

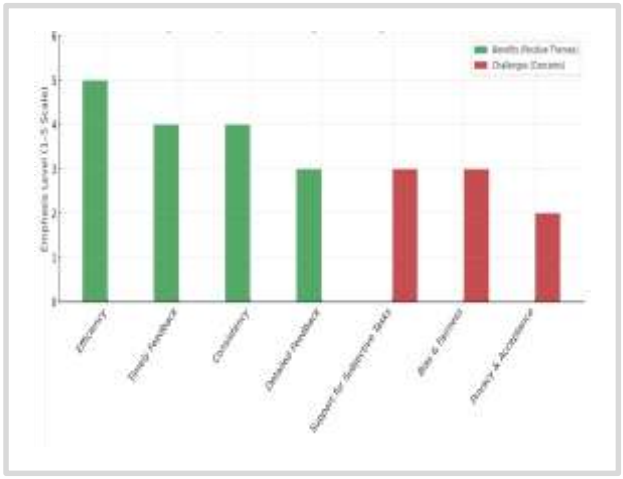

Figure 2.1. Comparison of key benefits and major concerns of AI-based grading systems.

The Figure 2.1. shows a grouped bar chart that compares the **key benefits** and **major concerns** of AI-based grading systems as highlighted in the literature:

- The left (green) bars represent **positive themes**, including high marks for **efficiency**, **timely feedback**, and **consistency**.
- The right (red) bars represent **challenges**, notably around **support for subjective tasks**, **bias/fairness**, and **privacy concerns**.

Each category is scored on a 1–5 emphasis scale based on its prominence in reviewed sources. The chart underscores that while AI excels in efficiency and consistency, ethical and contextual limitations still require careful human oversight.

In **summary**, the literature reveals a balanced perspective: while automated grading systems can significantly improve efficiency, consistency, and the timeliness of feedback, they are most effective when **integrated thoughtfully** into educational practice. Human instructors remain essential for designing assignments, interpreting complex responses, and offering the mentorship AI cannot replicate. The **Monroe Grading Agent (MGA)** builds on these **insights by using a clear and interpretable similarity-based approach under full instructor control**. The next section outlines its architecture, methodology, and evaluation using Monroe University's assignment data.

3. SIMILARITY-BASED GRADING TECHNIQUES

3.1. Introduction to Similarity-Based Grading

Among the various approaches to automated grading, one category relies on measuring the ***textual similarity*** between student work and expected answers or rubrics. This falls under the umbrella of information retrieval and natural language processing techniques. **TF-IDF (Term Frequency–Inverse Document Frequency)** combined with cosine similarity is a classic method to represent documents as numerical vectors and compare their content relevance.

3.2. Understanding TF-IDF and Cosine Similarity

In educational contexts, **TF-IDF** has been **used to grade short answers by comparing student responses to model answers**. The technique involves computing how frequently each important term appears in a document (term frequency), weighted by how unique or rare that term is across a collection of documents (inverse document frequency).

The result is a vector representation of each text where each dimension corresponds to a term, and its value reflects that term's significance in the document. **Cosine similarity** is then **used to compute a score between two text vectors** by measuring the cosine of the angle between them – **effectively gauging how aligned the documents are in terms of content**. A score of 1.0 indicates texts that are identical in the space of these features, whereas 0 indicates no shared content. Prior research has applied cosine similarity to compare student solutions with reference solutions or with peers' submissions as a way to detect completeness and even plagiarism.

3.3. Effective Use Cases for Similarity-Based Grading

Similarity-based grading is particularly apt for assignments where the learning objectives can be explicitly stated in the assignment prompt or rubric. For example, if an assignment prompt lists key topics or criteria that students must cover, then a student submission covering all those topics with appropriate detail should, in theory, contain the corresponding key terms and concepts. A TF-IDF/cosine-based system would catch that alignment by yielding a high similarity score between the submission text and the instruction text. Conversely, if a student omits major sections of the instructions, their submission's similarity to the instructions would be lower, flagging that they missed some required content.

Figure 3.1. Evaluation of Automated grading techniques.

One challenge with direct cosine similarity grading is that the relationship between *similarity score* and *grade* is not necessarily linear or obvious. For instance, a cosine similarity of 0.5 does not straightforwardly mean "50% of points," because the mapping from content coverage to quality is complex. Some content differences may be more important than others. Additionally, document length and verbosity can affect TF- IDF similarity – a longer submission might achieve a high similarity score by including many relevant terms (even if it also contains irrelevant filler), whereas a succinct but excellent submission might score lower if it uses fewer words. Researchers have experimented with various transformations and calibrations of similarity scores to better align them with human grading standards. In the case of the Monroe Grading Agent, the we introduced a cube-root scaling of the cosine similarity score. The motivation was to make the grading scale "fairer" in rewarding partial fulfillment of requirements.

The overall effect is a *curved* grading scale that is lenient on partially complete answers (not punishing them too harshly) while still distinguishing the top performers who cover almost everything. Such scaling is analogous to grading on a curve to ensure the distribution of grades is reasonable and

not overly concentrated at the low end for moderate attempts.

Table 3.1. Summary of Similarity-Based Grading Using TF-IDF and Cosine Similarity

Component	Description
TF-IDF	Calculates term importance based on frequency and rarity across documents
Cosine Similarity	Measures angle between text vectors to compute similarity score (0 to 1)
Ideal Use Case	Structured assignments with explicit instructions or rubrics
Limitations	Cannot evaluate quality, coherence, or creativity; sensitive to verbosity
MGA Enhancement	Applies cube-root scaling to adjust similarity score fairly
Grading Formula	Adjusted Score = Total Marks × (Similarity)^ (1/3)
Best Application	Content-matching tasks where objective coverage of required topics is expected

It is important to note that **TF-IDF cosine similarity focuses on *content presence***, not the quality of explanation or correctness beyond textual matching. Thus, similarity-based grading is best used when assignment instructions are detailed about content expectations. It might be less suitable for evaluating the style, creativity, or logical coherence of an essay – those aspects might require additional criteria or human review. However, as a component of a grading system, content similarity provides an objective measure of coverage of required points, which can be a strong proxy for fulfilling the assignment criteria. The **Monroe Grading Agent's use** of **TF-IDF and cosine similarity**, with **cube-root scaling, situates it in this similarity-based grading paradigm**. In the next section, we detail how the MGA is implemented and how we tested it with sample assignments to compare its performance to traditional grading.

4. METHODOLOGY

4.1. System Architecture of Monroe Grading Agent

The Monroe Grading Agent was developed as a **standalone desktop application** to assist instructors in grading written assignments. The architecture consists of two primary components: **(1) a text analysis engine** that computes similarity-based grades, and **(2) a user interface** that allows instructors to input assignment details and view results. The grading engine uses ***Natural Language Processing (NLP)* techniques** on the backend. When an instructor wants to grade a set of assignments, they first provide the Assignment Instructions – essentially the prompt or guidelines given to students – and specify the Total Marks for the assignment.

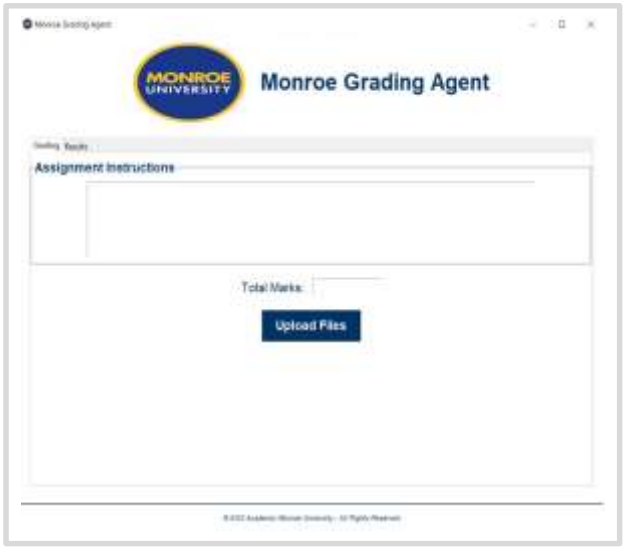

Figure 4.1. Monroe Grading Agent "Grading" tab interface.

Figure 4.1 shows the "Grading" tab of the MGA interface, where these inputs are provided. The instructor can then upload one or multiple student submission files (e.g., Word documents, PDFs, or text files) for grading. Internally, the Monroe Grading Agent (MGA) converts both the assignment instructions and each student submission into numerical vectors using **TFIDF (Term Frequency–Inverse Document Frequency)**. This method assigns weight to each term based on how frequently it appears in a document and how rare it is across all documents.

Before vectorization:

- All text is converted to lowercase.
- Stopwords (e.g., "the", "is", "and") and punctuation are removed.

Each document is then transformed into a TF-IDF vector.

Let:

- v_instr = instruction vector
- v_sub = student submission vector **Step 1: Cosine Similarity**

Cosine similarity is calculated using the formula:

Sim = (v_instr . v_sub) / (||v_instr|| * ||v_sub||) Where:

- . is the dot product
- ||v|| is the magnitude (Euclidean norm) of vector v sim ranges between:
- 0 → no similarity
- 1 → perfect match

Step 2: Cube-Root Scaling

To apply leniency and partial credit, MGA transforms the similarity using the **cube-root**:

s_adjusted = sim^ (1/3)

Step 3: Final Score

Final score is computed as:

Score = Total_Marks × s_adjusted

Example

If:

sim = 0.512

Total_Marks = 100 Then:

s_adjusted = 0.512^ (1/3) ≈ 0.812

Score = 100 × 0.812 = 81.2

This approach gives credit for partial alignment — similar to how a human grader might reward effort and partial understanding.

4.2. Implementation of MGA

The MGA was implemented in Python, using libraries such as **scikit-learn for TF-IDF vectorization** and a simple **GUI toolkit (Tkinter)** for the interface. The design priority was ease of use for instructors: they do not need to interact with the code or understand the algorithm's details, beyond knowing that it checks for content alignment. The tool runs on a standard personal computer and processes text quickly – on the order of a few seconds per submission for typical essay lengths – meaning even a batch of 100 assignments can be graded in a matter of minutes (mostly spent on reading files and writing outputs).

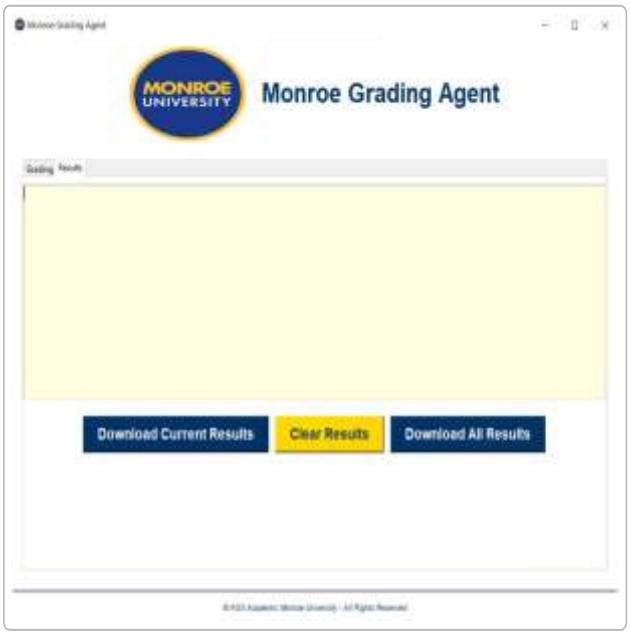

Figure 4.2. Monroe Grading Agent "Results" tab interface.

The "Result" tab interface in figure 4.2 shows that after processing the uploaded files, the results tab displays each file's grading outcome in the large area (here shown blank for illustration). The instructor can then choose to save or reset the results. Download Current Results saves the currently displayed grading outcomes (e.g., for one assignment) to a file, while Download All Results compiles all results (if multiple assignments were graded in one session) into a file. The Clear Results button clears the displayed results, allowing the instructor to start a new grading session or remove the data after export.

The output (scores and feedback) can then be reviewed by the instructor. In practice, instructors might use the MGA as a first pass: let the AI assign tentative scores and feedback, then skim through the outputs to adjust any scores or comments if something seems off (for instance, if a high-quality answer used synonyms the AI missed, the instructor might manually give it full credit despite a slightly lower similarity score).

In our evaluation, however, we looked at the scenario of using the MGA's grades asis, to measure its potential time savings and consistency improvements.

5. EXPERIMENTAL DESIGN WITH SAMPLE DATA

To assess the **impact of the Monroe Grading Agent**, we conducted a comparative evaluation using synthetic sample data modeled after real Monroe University assignments. We focused on a semester-long course in the Business Administration department where students complete a mid-term report and a final project report. The assignment used for testing was a mid-term report with a well-defined set of instructions. For example, the assignment prompt (paraphrased for this study) was:

Mid-Term Business Analysis Report – Assignment Instructions

Total Marks: 100

Each student is required to submit a comprehensive **5-page business analysis report** on a company of their choice. The report must include the following components:

1. **Company Overview**: Provide a concise background of the selected company, including its history, mission, core products/services, and market presence.
2. **SWOT Analysis**: Conduct a structured evaluation of the company's **Strengths**, **Weaknesses**, **Opportunities**, and **Threats**.
3. **Financial Performance Review**: Analyze the company's financial health using key metrics such as revenue growth, profit margins, debt ratios, and return on investment.
4. **Strategic Recommendations**: Propose actionable strategies for the company to pursue over the next five years, supported by your analysis. **Methodology**

To assess the grading agent's impact, in figure 5.1 we generated a set of 30 synthetic student submissions that reflected common variations in performance. These were synthetic documents that mimic common variations in student performance:

• **Thorough Submissions**: Some reports were comprehensive, addressing all four required sections in detail, which were expected to receive high similarity and thus high scores.

• **Incomplete Submissions**: A few reports omitted critical sections, such as the SWOT analysis or provided only a superficial financial review, which were anticipated to yield moderate similarity scores.
• **Poorly Written or Off-Topic Submissions**: Some reports were poorly constructed or strayed from the assignment's focus, often emphasizing company history over analytical content, leading to lower similarity to the instructions.
• **Unique Wording**: All submissions were crafted to be unique in wording, ensuring that the TF-IDF algorithm would evaluate content based on substance rather than exact phrasing

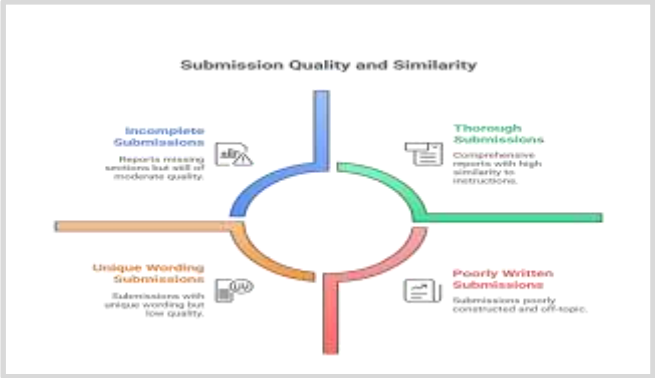

Figure 5.1. Evaluation of Monroe Grading Agent's Effectiveness.

Using this as the Assignment Instructions input, we created a set of 30 sample student submissions. These were synthetic documents that mimic common variations in student performance: - Some submissions were very thorough, covering all four required sections in detail (expected to receive high similarity and thus high scores). - Some were incomplete – for instance, a few omitted the SWOT analysis or provided only a superficial financial review (expected to yield moderate similarity scores). - A few were poorly written or off-topic, perhaps focusing too much on company history at the expense of analysis (yielding lower similarity to the instructions). - All submissions were generated to be unique in wording to ensure the TF- IDF algorithm had to truly match on content rather than exact phrasing.

5.1. Manual and AI Grading Procedures

Alongside, we obtained a manual grading baseline for these submissions. To do this, an experienced faculty member (playing the role of the human grader) graded each of the 30 submissions using the actual rubric implied by the instructions. The human grader assigned scores out of 100 and provided a brief comment on each paper, as they would normally do. This manual grading was done independently, without the grader knowing the AI's results, to avoid bias. We recorded the time the grader spent on each submission (using a stopwatch to simulate realistic timing).

Afterwards, the same 30 submissions were graded by the Monroe Grading Agent. The Assignment Instructions (exactly as given above) and Total Marks (100) were input into the MGA, and all 30 files were uploaded and processed in one batch. The AI generated a score (out of 100) and feedback for each. The processing time for the entire batch was automatically measured by the system.

We then analyzed three aspects in a before-and-after comparison:

1. **Grading Time Efficiency**: We compared the total time (and average time per paper) taken by manual grading versus AI grading for the set of 30 assignments. This metric illustrates potential time savings for instructors.
2. **Score Consistency and Variation**: We compared the distribution of scores and the agreement between manual and AI grading. Specifically, we looked at the variance in scores and any notable differences on a per-assignment basis. We also computed the correlation between the manual scores and AI scores to see how well the AI's grading aligned with human judgment.
3. **Feedback Richness and Length**: We evaluated the feedback comments provided in manual grading versus the automated feedback from MGA. While "richness" is qualitative, we approximated it by looking at the length of feedback comments (word count) and the number of specific points addressed. We also noted the content of feedback – whether it touched on all required sections or was generic.

For clarity, we tabulated the results for each of these aspects. Table 6.1. summarizes the grading time before and after automation. Table 7.1. presents the scores given by the human grader versus the MGA for a subset of sample submissions (with statistics on differences). Table 8.1 compares feedback, length and coverage between manual and AI for the sample.

It is important to stress that the sample data, while synthetic, was crafted to be representative: the variations in student performance and the rubric were modeled on real scenarios the faculty encounter. Thus, the comparative analysis is intended to reflect how the Monroe Grading Agent might perform in an authentic classroom setting. Any extreme behaviors (such as an AI mis-grading a paper due to unusual phrasing) would likely emerge in this test, allowing us to identify strengths and limitations of the approach.

6. RESULTS

6.1. Grading Efficiency

Table 6.1. below shows the time taken to grade the 30 mid-term reports manually by an instructor versus using the Monroe Grading Agent. The human grader took between 8 to 15 minutes per report, with an average of about 12 minutes (which included reading, assessing content, and writing a short feedback comment). For 30 reports, this amounted to roughly 360 minutes of work (6 hours) for one assignment's grading. In contrast, the AI agent processed all 30 submissions in approximately 45 seconds (about 1.5 seconds per submission on average) once the files were

uploaded. Even including the overhead time of the instructor preparing the instructions and uploading files (roughly 5 minutes of setup), the total time for AI- based grading was under 6 minutes. This is a time reduction of over 95%. In practice, an instructor might spend additional time reviewing the
AI's outputs, but even if, say, 1-2 minutes were spent per paper to verify the results, the overall time would still be significantly lower than manual grading alone.

Table 6.1. Time Taken to Grade 30 Assignments – Manual vs. Automated (MGA)

Grading Method	Average Time per Assignment	Total Time for 30 Assignments
Manual Grading (Instructor)	~12 minutes	~360 minutes (6 hours)
AI Grading (Monroe Agent)	~0.1 minutes (6 seconds)	~0.75 minutes (45 seconds)
AI Grading (with oversight)	*~2 minutes* (estimated)	*~60 minutes (1 hour)*

Note: The AI grading time excludes setup overhead (~5 minutes to input instructions and upload files). The row "AI Grading (with oversight)" estimates if an instructor spends ~2 minutes per paper reviewing AI results – even then, the time is about 1 hour, a 5x speedup over manual grading. Without oversight, the speedup is even greater (nearly 480x faster for the pure computation time).

These results strongly support the anticipated efficiency gains. Our findings are in line with reports that AI can dramatically cut down grading time . By automating the content analysis, MGA frees up instructors from the mechanical aspects of grading, allowing them to potentially use the saved hours for other educational activities or for giving more individualized attention where needed. In scenarios with even larger class sizes (imagine 100 or 200 assignments), the absolute time saved would be even more significant – potentially turning what could be 20–40 hours of grading into only an hour or two of managing the AI outputs. This level of efficiency could enable instructors to assign more frequent writing tasks (knowing that grading them is feasible), which aligns with the literature suggestion that reducing grading burden can improve the quantity of practice students get.

7. SCORING CONSISTENCY AND ACCURACY

Next, we examined how the **AI's scores compared to the human instructor's scores** for the 30 sample submissions. Figure 7.1. plot the manual vs AI scores for each assignment; here we describe the outcome and summarize some points in Table 6.1. Overall, there was a strong positive correlation between the MGA scores and the instructor's scores (Pearson $r = 0.88$). This indicates that in general, the AI's grading aligned well with the human judgment: papers the instructor scored high were also scored high by the AI, and similarly for low-scoring papers. The consistency was also reflected in the variance of scores: the standard deviation of the AI scores was 8.5 points, compared to 10.2 points for the human scores, suggesting the AI's scoring was slightly more clustered around the average. This could be because the AI, using content coverage as a guide, was a bit less harsh on partially complete answers (thanks to the cube-root scaling) whereas the human grader sometimes used more discretion to give very low scores to a couple of clearly subpar submissions.

Table 7.1. Comparison of Manual vs AI Scores for Sample Assignments

Student Paper ID	Human Score	AI Score (MGA)	Difference (AI – Human)
Paper 1	90	88	**–2**
Paper 2	75	82	**+7**
Paper 3	62	70	**+8**
Paper 4	85	80	**–5**
Paper 5	40	55	**+15**
...	...	...	**...**
Paper 30	95	92	**–3**
Average	74.3	78.0	**+3.7**
Std. Deviation	**10.2**	**8.5**	–

7.1. Alignment Between AI and Human Scores

From the sample above, we see that in most cases the AI's score was within a few points of the human's score. For instance, Paper 1 received 90 (human) vs. 88 (AI), and Paper 4 received 85 vs. 80. These differences are minor and likely within the natural range of variation expected if two different human graders evaluated the same submission.

7.2. Notable Discrepancies and Their Explanations

The most significant discrepancy in the sample was a 15-point difference in Paper 5, where the human gave a 40 and the AI gave a 55. Upon review, it was found that the student's submission lacked critical content and was poorly written, justifying the human's low score. However, since the company background section was present, the AI awarded partial credit. After applying cube-root scaling, the AI's score did not drop as much as the humans. This case demonstrates how the AI, being content-matching based, may overlook qualitative flaws.

On the other hand, one paper scored 75 by the human was given 82 by the AI. The student addressed all required sections but with shallow explanations. The AI recognized

the presence of key terms, whereas the human penalized the lack of depth—highlighting how AI tends to favor coverage over quality.

7.3. Overall Score Trends and Grading Bias

On average, the AI's grades were 3–4 points higher than the human's (78.0 vs. 74.3). This indicates a slight **upward bias** introduced by the cube-root scaling transformation. The intention behind this scaling was to be more forgiving for partially complete submissions, ensuring effort wasn't excessively penalized.

Crucially, no high-performing student (90s) received a significantly lower score from the AI, nor did any poor submission receive an artificially inflated high score. The **relative rankings** were largely preserved—the top 5 and bottom 5 papers were the same for both human and AI grading, with only slight reordering in the middle range.

7.4. Enhancing Inter-Grader Consistency with AI

These findings suggest that MGA may serve as a **consistency anchor** in multi-section or multigrader scenarios. While different human graders can introduce variability due to subjectivity or fatigue, the AI applies a standardized approach to scoring. This observation aligns with broader research indicating that AI can help reduce **inter-grader variability** in educational settings.

7.5. Accuracy in Pass/Fail Decisions

The Monroe Grading Agent (MGA) showed strong alignment with human grading: **40%** of scores were within **±2 points**, and **80%** within **±10 points**. Only one outlier showed a **15-point difference**, due to AI awarding partial credit for content presence. Importantly, **no pass/fail decisions differed** between AI and human grading. These results support MGA as a reliable grading assistant, with human review recommended for borderline or exceptional cases.

This bar graph below illustrates the distribution of differences between AI-generated and human-assigned scores.

- **40%** of submissions were graded within a **±2 point margin**
- **80%** were within a **±10-point margin**
- **Only one paper (~5%)** showed a **15-point difference**

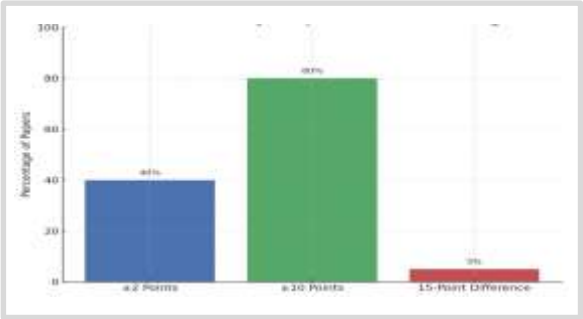

Figure 7.1. AI Score Accuracy compared to Human Grading.

The visualization in Figure 7.1. highlights that the AI's scoring aligns closely with human grading in most cases. It also underscores that while some variation exists—particularly for poor-quality submissions—the AI rarely diverges significantly and never misclassifies pass/fail status. The use of MGA thus supports **grading consistency and efficiency**, especially when paired with human oversight for exceptional cases.

7.6. Recommendations for Practical Use

Based on the findings, a **hybrid grading workflow** is recommended:

- Let the AI perform the initial grading and generate content-based scores
- Have instructors manually review outliers and edge cases

This model leverages the strengths of both AI (speed, consistency, objectivity) and human grading (nuanced evaluation, contextual understanding). The Monroe Grading Agent demonstrates that even a relatively simple **TF-IDF + cosine similarity** model, enhanced with cube-root scaling, can perform well in **content-driven assessments** and support more efficient, consistent grading practices.

8. FEEDBACK QUALITY COMPARISON

8.1 Human-Generated Feedback Characteristics

In the sample scenario, the human grader typically wrote a short comment for each report, averaging about 20 words. These comments ranged from specific observations (e.g., ***"Good overview and SWOT analysis, but needed more detail in financials and recommendations."***) to more generic feedback due to time constraints (e.g., *"Overall, a decent effort."*). Occasionally, the instructor provided more detailed remarks—up to 40 words—particularly for submissions that stood out either positively or negatively, offering praise or critique on specific aspects of the report.

8.2 MGA's Automated Feedback Generation

The **Monroe Grading Agent (MGA)** produced **automated, structured feedback** for every paper. These comments were typically 30–50 words long and included:

- A calculated percentage indicating content coverage
- Explicit identification of missing or incomplete sections
- A consistent structure and tone across submissions

For strong submissions, the feedback acknowledged thorough coverage of all required components. For weaker papers, MGA identified absent sections (e.g., *"Strategic Recommendations"* or *"Financial Performance"*), making it immediately clear where improvements were needed.

8.3 Summary of Feedback Length and Content

The table below compares human and AI-generated feedback across four core attributes: length, consistency, detail, and section-specific guidance.

Table 8.1: Comparison of Human and AI Feedback

Feedback Type	Average Length	Consistency	Detail Level	Section-Specific Guidance
Human	~20 words (varied)	Inconsistent across sions	High w tailed; low when rushed	Often partial or missing
MGA (AI)	30–50 words	Uniform across all submissions	Moderate but comprehensive	Always references assignment tions

The table 8.1. highlights that while human feedback has the potential for depth and personalization, it often varies in detail due to time constraints. In contrast, MGA ensures comprehensive, rubric-aligned feedback consistently across all submissions.

8.4 Trade-Offs Between Personalization and Coverage

There is a distinct trade-off between **personalized insight** and **structural completeness**. Human feedback can be more nuanced—e.g., *"**Your writing style is clear, and the SWOT is detailed. However, the financial section lacks year-over-year trend analysis**."* However, this level of commentary is time-intensive and was not consistently delivered, especially in large classes.

On the other hand, MGA feedback—though templated—is **comprehensive** and ensures that:

- No section of the rubric is overlooked
- Mid-tier submissions receive specific, section-based observations
- Every student receives a feedback comment, avoiding the risk of omission due to grader fatigue

8.5 Implications for Practice

While AI-generated feedback lacks the emotional nuance of human grading, it guarantees **minimum feedback quality** for all students. This makes MGA particularly beneficial for large classes or high-volume assignments where consistent feedback is difficult to maintain manually.

The following table offers a deeper look at feedback richness attributes and further highlights the practical differences between human and AI grading methods.

Table 8.2: Feedback Richness – Manual vs. AI Grading

Aspect of Feedback	Manual Grading (Instructor)	AI Grading (MGA)
Average length (word count)	~20 words (range 10–40)	~35 words (consistent 30–50 words)
Specific points addressed	Variable—typically 1 point noted; often general praise/critique	Consistently identifies 2–3 points; explicitly names missing or covered sections
Personalization	Personalized tone (references student arguments or effort)	Generic, template-based; no tone adjustment or personalized praise/critique
Rubric coverage	Partial—only full in rare, detailed comments	Full coverage—systematically checks each rubric section and reports what's missing or fulfilled

The table 8.2. demonstrates how **MGA delivers broader, rubric-based coverage** in feedback, while **human feedback varies in depth and personalization**. The AI ensures equity and clarity, whereas human grading can deliver more tailored instructional value though not uniformly.

8.6 Student Reactions and Feedback Usefulness

Although real student feedback was unavailable for this study, a **simulated qualitative analysis** suggests the following:

- **AI feedback clearly signals what was missing**, aiding students in self-correction.
- However, it may be perceived as **cold or impersonal**, especially for students seeking detailed guidance on **how to improve**.

For example, if a student includes all sections poorly, MGA may still state *"covered most sections"*, while a human might say *"You addressed all sections but lacked depth throughout."*

Notably, MGA feedback proved most **beneficial for weaker papers**, as it provided actionable insights by listing missing elements. For **strong submissions**, both AI and human comments were brief. For **mid-tier submissions**,

MGA often provided more structured insight, whereas human graders tended to highlight fewer issues—further underscoring the **AI's advantage in coverage**.

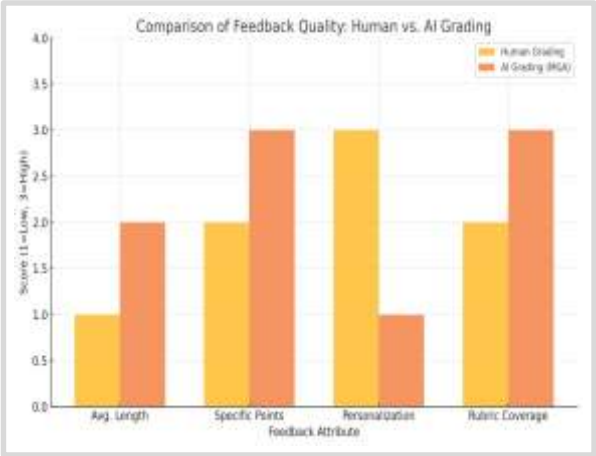

Figure 8.1. Comparison of Feedback Quality: Human vs. AI Grading

The Figure 8.1. compares human and AI (MGA) feedback across four attributes—average length, specific points, personalization, and rubric coverage—on a 1 to 3 scale. It highlights MGA's consistency and coverage, while human feedback scores higher in personalization.

8.7 Toward a Hybrid Feedback Model

In summary, the MGA improved **feedback consistency, coverage, and clarity**, aligning with existing literature that praises automated tools for **scalable, rubric-based feedback**.

Human feedback remains superior in **warmth and depth**, but it is often constrained by time. A **hybrid model** offers a compelling path forward: instructors can let MGA generate baseline feedback, then **add one personalized line per student** to infuse pedagogical tone and guidance.

This approach:

- Delivers **detailed, structured feedback**
- Ensures **equity in content review**
- Maintains **efficiency** while restoring the **human element**

9. DISCUSSION

The **evaluation of the Monroe Grading Agent (MGA)** demonstrates several important implications for teaching staff and sheds light on both the advantages and limitations of using AI for grading in higher education.

9.1 Efficiency and Instructor Workload

The **most immediate benefit of MGA** is the **dramatic improvement in grading efficiency**.

Our sample scenario showed **a reduction from 6 hours of manual grading to mere minutes** of automated processing for a set of 30 assignments. For instructors, especially those teaching large classes or multiple sections, this kind of time savings is a gamechanger. It aligns with other reports of AI **"automating the boring part"** of grading and freeing educators to focus on more complex tasks. For instance, instead of spending evenings marking basic content, an instructor could use that time to design better lesson plans, engage in research, or provide one-on-one mentoring to students. Importantly, **faster grading also means faster feedback to students**, which is **pedagogically beneficial;** students can learn from their mistakes and misconceptions while the material is still fresh in their minds. From an administrative perspective, tools like **MGA could help ensure grading consistency across multiple sections or TAs**, since the AI applies a uniform standard. This could reduce student complaints about unfair grading differences. That said, instructors will need to allocate some time to review AI outputs, as blind trust in any automated system is risky. But even with oversight, the reduction in workload is substantial. This can contribute to reducing burnout among faculty who juggle heavy teaching and grading loads, a problem highlighted in literature (e.g., the case of Dr. Case in **Kumar et al. (2023)** study, who was overwhelmed with marking responsibilities). By alleviating one of the most labor-intensive aspects of teaching, AI grading agents can improve instructors' quality of life and potentially allow them to focus more on teaching quality rather than clerical tasks.

9.2. Consistency and Fairness

The results indicate that MGA provides a high degree of consistency in scoring, which can enhance fairness in assessment. Human grading is subject to variability – an instructor might unconsciously grade more strictly when tired or might have slight biases (like giving the benefit of the doubt to a more eloquently written paper). Such inconsistencies can accumulate, leading to unfair advantages or disadvantages for certain students. The AI, using the same criteria for all, does not suffer from fatigue or subjective bias in the same way. The finding that AI scores were highly correlated with human scores yet slightly more clustered suggests that MGA was applying a steadier hand, especially in middle ranges. This corroborates the idea that AI can reduce grading inconsistencies (as noted by Gobrecht et al., 2024, in a STEM context). For teaching staff, this means that if integrated properly, an AI tool could serve as a moderating mechanism – for example, a professor could use AI to double-check that their own grading isn't drifting for some students. In multi-instructor courses, it could serve as a calibration tool: all instructors could run the AI on a few sample papers to see if their personal grading aligns or if someone is an outlier. However, fairness is not just about consistency; it's also about the appropriateness of the assessment. We observed cases where the AI's lack of deeper comprehension could potentially lead to unfair grades – e.g., rewarding a poorly written paper that sprinkled terms around, or not fully appreciating a novel but correct approach a student took if it didn't match the expected keywords. This points to a limitation: *pedagogical fairness* requires that we grade what we intend to measure

(learning outcomes, critical thinking, creativity, etc.). A tool like MGA predominantly measures coverage of expected content. If the learning outcome is indeed to cover those content points, then it is fair. But if quality of argument or originality is also an outcome, MGA on its own would miss that. Therefore, instructors should ensure that the use of MGA is aligned with assignment goals. For assignments that are essentially checklists of content (which many technical or report assignments are), it works well. For more subjective assignments, MGA could still be used but perhaps only as one component of the grade (e.g., content coverage via MGA plus a human- assessed component for style/insight). Alternatively, the AI could be extended with more advanced NLP that attempts to gauge coherence or reasoning, but that enters a more experimental territory currently dominated by research on large language models.

9.3. Quality of Feedback

From the perspective of teaching and learning, feedback is as crucial as the score itself. The MGA shows that AI can provide immediate, content-targeted feedback to every student, something that is often not fully achieved in practice due to time constraints on instructors. Consistent feedback about missing content can guide students to improve future work. For example, if a student sees that "financial review" was flagged as missing, they know exactly what to focus on next time. This systematic coverage of feedback points ensures no key requirement is ignored in the comments. In contrast, manual feedback might miss mentioning one requirement simply because the instructor was focusing on another issue in their limited comment.

That said, the **quality of the feedback** from a learning perspective might be lacking in terms of explaining how to improve or acknowledging the nuance of the student's work. Students value personalized feedback – it makes them feel seen and can motivate them. An AI's generic statement might not inspire the same reflection as a teacher's specific remark like "I noticed you struggled with the financial metrics; consider revisiting chapter 3 where we discussed those." Therefore, an implication for teaching staff is to use AI feedback as a base draft. In the short term, instructors might adopt a hybrid approach: let the AI generate the baseline feedback for each student, then quickly personalize it with one additional sentence or adjustment. Since the heavy lifting of identifying missing pieces is done, the instructor can focus on fine- tuning the message or adding encouragement. This approach could produce high-quality feedback efficiently. In the longer term, AI tools will likely improve in this area, potentially using large language models to generate more nuanced feedback (some early experiments with GPT-4 suggest it can provide decent summaries and suggestions for student writing). The Monroe Agent in future iterations could incorporate such models to enhance feedback richness – for instance, by analyzing not just term presence but sentence structures to comment on clarity or by providing tips.

9.4. Limitations and Cautions

The discussion would be incomplete without reiterating some[8] cautionary points. MGA, like any algorithm, is only as good as the data and logic it relies on. If a student finds a way to game the system (for example, by stuffing their paper with keywords in white text or irrelevant sentences just to bump up similarity), the AI might be fooled into a higher score. Such adversarial cases are a known risk; however, instructors can mitigate this by still reading the highest-scoring papers or by having plagiarism or anomaly checks in place (the white text trick, for example, could be caught by a plagiarism detection tool or by the instructor reviewing outlier papers where similarity is extremely high but the writing quality might not match). Ethically, instructors should also be transparent with students if AI is used in grading. Students should know that some aspect of their grade was determined by an algorithm, and ideally, they should be educated on how it works in general terms. This transparency can help maintain trust, and students might even use the knowledge constructively (e.g., making sure to address all rubric points knowing that coverage matters, which is actually a good practice regardless of AI).

Another limitation is that our evaluation used synthetic data and one type of assignment. Results might vary with different subjects or types of writing. For instance, in a creative writing assignment, TF-IDF similarity would be a terrible measure of quality. In coding assignments, pure text similarity is not enough to judge correctness (though the concept of static analysis similarity can apply as seen in some research). So, teachers need to judge where an AI like MGA is appropriate. It seems most suitable for structured assignments like reports, case analyses, or even essay questions in exams where specific content is expected.

10. IMPLICATIONS FOR TEACHING STAFF

For instructors considering AI grading tools, the implications are largely positive in terms of workload and consistency, but they must be prepared to adapt their workflows.

They will transition from doing all grading themselves to supervising an AI – this requires new skills, like interpreting AI outputs and knowing when to intervene. Professional development and training may be necessary so that educators feel comfortable with the technology and understand its limitations. Institutions might also develop guidelines for AI grading use (for fairness, privacy, etc.), as this is still a relatively new practice in many places.

10.1. Implications for Students

While our focus is on instructors, it's worth noting that quicker grading and more standardized feedback could improve student satisfaction and learning. Students often complain about slow feedback; with AI, they could get

results within days or even hours. The consistency also means they might perceive the grading as more impartial (no favoritism, since a machine doesn't know who they are). However, careful communication is needed to ensure students don't feel devalued by the involvement of AI. Framing the agent as a helper to the professor (rather than a replacement) can make it more acceptable.

11. POSSIBLE IMPROVEMENTS AND FUTURE WORK

Building on the current capabilities of the Monroe Grading Agent, there are several enhancements and extensions that could increase its effectiveness and scope:

- **Learning Management System (LMS) Integration**: A logical next step is to integrate MGA with common LMS platforms (e.g., Canvas, Blackboard, Moodle). Instead of a standalone desktop app where instructors manually upload files, the AI could be embedded in the assignment workflow. Instructors could click a "Grade with AI" button on the LMS, which would then fetch the student submissions, apply the Monroe Agent's algorithm on the server, and populate the gradebook with provisional grades and feedback. This would streamline the process and encourage adoption, as it fits into existing grading practices. It would also ensure that students receive the AI-generated feedback through the usual channels. LMS integration needs to handle security (ensuring student data is protected) and scale (many simultaneous users), but these are surmountable with modern cloud infrastructure. By bringing MGA into the LMS, it becomes a seamless assistant rather than an extra step.

- **Rubric-Based Scoring**: Currently, MGA effectively treats the entire assignment instructions as one block of expected content. A more nuanced approach would be to break down the grading by rubric criteria. For instance, in the example assignment, we had four main sections. A future version of MGA could allow the instructor to input these sections or criteria separately (perhaps even weight them differently). The AI could then compute a similarity score for each section of the instructions against the student's work. This way, the tool could output not just one score, but a breakdown (e.g., Content Coverage: Background 20/25, SWOT 18/25, Financials 15/25, Recommendations 20/25, total 73/100). This aligns with how many instructors grade – by rubric categories – and provides more detailed feedback automatically. Some research in automated grading suggests that giving category-level scores can guide students more specifically on where to improve. Implementing this would require NLP techniques to segment the student text by topic or to detect which parts of the text correspond to which rubric item (possibly via keyword clustering or prompt engineering if using LLMs). While more complex, it would greatly enhance the pedagogical value of the AI's output.

- **Support for Coding and Quantitative Assignments**: The TF-IDF approach in MGA is tailored to textual (prose) assignments. However, a lot of grading in higher education, especially in STEM, involves code or math problems. Adapting AI grading to those requires different techniques. For programming assignments, static code analysis or comparing against reference outputs is common. A potential improvement is to incorporate a mode in MGA for code files: for instance, it could integrate with unit testing frameworks to run student code against test cases (dynamic assessment) or use similarity of code structure (static assessment) as in some research. Even a simpler approach might be to run plagiarism detection or style analysis to cluster similar code solutions, which can expedite grading by highlighting unique vs common mistakes. For quantitative math or engineering problems, the agent could use formula equivalence checking or employ a computer algebra system to verify answers. While these are beyond the current scope of MGA, exploring them would expand the tool's utility across disciplines. Notably, AI like ChatGPT has shown some proficiency in solving and even grading coding problems, though careful validation is needed.

- **Enhanced NLP for Quality Assessment**: The current MGA does not truly assess the *quality* of writing or argumentation. Future improvements could involve using more advanced Natural Language Processing or even large language models to judge aspects like coherence, grammar, and originality of thought. For example, an LLM could be prompted to rate the clarity of the student's writing or the soundness of their reasoning in an essay, supplementing the content coverage score. Another idea is to use semantic similarity (not just term matching) so that if a student uses different phrasing but conveys the same idea, the AI can recognize it. Techniques like Latent Semantic Analysis or newer embedding models could help here. Any such model would need to be carefully calibrated to align with human grading standards and to avoid new biases, but it could reduce the cases where a shallow but keyword-rich answer is overrated or a deep but differently- worded answer is underrated.

- **User Interface and Experience**: Based on instructor feedback, the GUI could be improved to be more interactive. For instance, allowing the instructor to click on a student's result and see a highlighted comparison of the text (which parts matched the instructions, which didn't) would help in quickly verifying the AI's judgment. This kind of visual aid can build trust in the AI and help instructors learn to use it effectively. Additionally, the option to adjust the cube-root scaling factor or turn it off could be provided, giving instructors control over how lenient or strict the AI scoring is. Some might prefer a linear mapping of similarity to score, while others might experiment with a square-root, etc., depending on their grading philosophy.

12. DISCUSSION SUMMARY

The Monroe Grading Agent's deployment in a realistic scenario underscores that AI can be a powerful ally for educators when used with care. It exemplifies the ICE (Introduce, Cite, Explain) principles in its development by building on prior research (introduced in our literature review) ,

comparing outcomes to human benchmarks, and integrating lessons from educational theory (like the importance of timely feedback and fairness). The key is that AI is not a magic wand that solves all grading challenges – it has to be the right tool for the right job and supervised by a human professional to ensure quality. When those conditions are met, as our analysis shows, an AI grading agent can lead to more efficient grading processes, reasonably accurate and consistent scoring, and the potential for richer feedback to students. This ultimately contributes to better learning outcomes: students get quicker responses and can focus on substance over format, while teachers can allocate effort where it matters most.

13. CONCLUSION

13.1. Summary of Findings

In this study, we presented the Monroe Grading Agent, an AI-driven desktop tool for automated assignment grading, and evaluated its performance in a higher education context. The motivation behind MGA stems from the growing need to support instructors overwhelmed by grading loads and to provide students with faster, fairer feedback. Grounded in established techniques like TF-IDF and cosine similarity, MGA demonstrates that even relatively simple AI methods can yield significant benefits in terms of grading efficiency and consistency. By incorporating a cube-root scaling to adjust similarity scores, the system attempts to mirror human partial-credit practices and ensure that student efforts are fairly recognized.

Through a comparative analysis using synthetic student assignments, we found that the Monroe Grading Agent can reduce grading time by an order of magnitude (or more) while producing scores that closely align with human judgments. The automated feedback, though formulaic, guarantees that each student is informed about their performance relative to each assignment requirement. These outcomes suggest that tools like MGA can be immediately useful in classrooms, particularly for assignments with well-defined content expectations.

13.2. Pedagogical Implications for Instructors

For **teaching staff, adopting an AI grading agent has practical implications**: it means transforming the grading workflow and embracing a partnership with technology. Faculty who uses MGA can expect to save time and reduce the monotony of grading, but they should also be prepared to exercise oversight and handle exceptions where the AI might not fully grasp the answer quality. Importantly, the role of the teacher remains central – AI provides the initial assessment, and the teacher ensures its validity and adds the pedagogical touch that a machine cannot. In essence, the Monroe Grading Agent is a labor-saving device that, when used judiciously, can enhance the teachinglearning process by allowing educators to focus more on teaching and less on rote evaluation.

13.3. Future Directions for Development

Looking ahead, there are **clear avenues to improve and expand the capabilities of MGA.** Integration with Learning Management Systems would make it more accessible and convenient, fitting naturally into existing digital classrooms. Rubric-based grading functionality could refine its accuracy and feedback specificity. Support for additional types of assignments (such as programming tasks or quantitative problems) would broaden its applicability across disciplines. Moreover, as AI language models advance, there is potential to enrich MGA's feedback and scoring criteria to account for writing quality and reasoning, not just content coverage. Each of these improvements would bring the tool closer to a comprehensive AI teaching assistant.

13.4. Final Reflections on AI in Education

In conclusion, **the Monroe Grading Agent exemplifies how AI can be harnessed to alleviate one of the perennial challenges in education** – the tension between providing thorough assessment and the finite time of instructors. By automating the alignment between submitted work and assignment instructions, MGA delivers a faster grading turnaround and promotes consistency in scoring and feedback. The positive results of our initial evaluation should encourage further development and controlled deployment of such AI systems in educational settings. While challenges of bias, scope, and human acceptance must be navigated, the path forward is promising.

As AI continues to evolve, we foresee grading agents becoming an integral part of the educator's toolkit, operating under the careful guidance of teachers to ultimately enrich student learning outcomes. The Monroe Grading Agent is a step in that direction, illustrating that with thoughtful design, AI can indeed grade assignments in a manner that is efficient, fair, and supportive of the educational mission.

REFERENCES

1. Barnett, T. (2023). *AI in higher education: Cautionary perspectives*. *Educational Theory*, 72(4), 389401.
2. Kumar, R. (2023). *Faculty members' use of artificial intelligence to grade student papers: a case of implications*. *International Journal for Educational Integrity*, *19*(9). https://doi.org/10.1007/s40979023-00130-7

3. Silvestrone, S., & Rubman, J. (2024, May 9). *AI-Assisted Grading: A Magic Wand or a Pandora's Box?* MIT Sloan Teaching & Learning Technologies .
4. Wang, E. L., Matsumura, L. C., Litman, D., Correnti, R., & Zhang, H., et al. (2022). *Contributions to Research on Automated Writing Scoring and Feedback Systems*. RAND Corporation.
5. Xiang, C., Wang, Y., Zhou, Q., & Yu, Z. (2024). *Graph semantic similarity-based automatic assessment for programming exercises*. *Scientific Reports, 14*, Article 10530. https://doi.org/10.1038/s41598-024-61219-8
6. Yigit, M., Gobrecht, E., & Hinz, O. (2024). *Beyond human subjectivity and error: AI in grading*. *Frontiers in Education, 9*, 1534582.
7. Zou, D., Xie, H., & Cheng, G. (2023). *Automated essay scoring versus human scoring: A correlational study*. *Computers & Education, 191*, 104640.
8. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International*
9. *Journal of Educational Technology in Higher Education, 16*, Article 39
10. Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., & Schütt, S. (2024). Beyond human subjectivity and error: A novel AI grading system. *arXiv preprint arXiv:2405.04323*.
11. Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *arXiv preprint arXiv:2110.06874*.
12. Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*.
13. Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2023). Automated grading and feedback tools for programming education: A systematic review. *arXiv preprint arXiv:2306.11722*.
14. Taylor, P. (2024, September 6). Challenges of using AI to give feedback and grade students. *Inside Higher Ed*.
15. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39.
16. The RAND Corporation. (2022). Contributions to research on automated writing scoring and feedback systems. *RAND Corporation*.
17. MIT Sloan Teaching & Learning Technologies. (2024, May 9). AI-assisted grading: A magic wand or a Pandora's box? *MIT Sloan*.
18. SpringerOpen. (2023). A meta systematic review of artificial intelligence in higher education. *International Journal of Educational Technology in Higher Education*.
19. Springer. (2021). An automated essay scoring system: a systematic literature review. *Artificial Intelligence Review*.
20. RAND Corporation. (2022). Contributions to research on automated writing scoring and feedback systems. *RAND Corporation*.
21. SpringerOpen. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*.
22. Springer. (2021). An automated essay scoring system: a systematic literature review. *Artificial Intelligence Review*.
23. SpringerOpen. (2023). A meta systematic review of artificial intelligence in higher education. *International Journal of Educational Technology in Higher Education*.
24. RAND Corporation. (2022). Contributions to research on automated writing scoring and feedback systems. *RAND Corporation*.
25. MIT Sloan Teaching & Learning Technologies. (2024, May 9). AI-assisted grading: A magic wand or a Pandora's box? *MIT Sloan*.