A User-Centered Design Approach to Evaluating the Usability of Automated Essay Scoring Systems

Erin E. Hall

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Mohammed Seyam, Chair
Dan Dunlap
Danfeng (Daphne) Yao

August 10th 2023
Blacksburg, Virginia

Keywords: Usability, Algorithmic Transparency, Machine Learning, AI explainability, Writing, Feedback

A User-Centered Design Approach to Evaluating the Usability of Automated Essay Scoring Systems

Erin E. Hall

(ABSTRACT)

In recent years, rapid advancements in computer science, including increased capabilities of machine learning models like Large Language Models (LLMs) and the accessibility of large datasets, have facilitated the widespread adoption of AI technology, such as ChatGPT, underscoring the need to design and evaluate these technologies with ethical considerations for their impact on students and teachers. Specifically, the rise of Automated Essay Scoring (AES) platforms have made it possible to provide real-time feedback and grades for student essays. Despite the increasing development and use of AES platforms, limited research has specifically focused on AI explainability and algorithm transparency and their influence on the usability of these platforms. To address this gap, we conducted a qualitative study on an AI-based essay writing and grading platform, with a primary focus to explore the experiences of students and graders. The study aimed to explore the usability aspects related to explainability and transparency and their implications for computer science education. Participants took part in surveys, semi-structured interviews, and a focus group. The findings reveal important considerations for evaluating AES systems, including the clarity of feedback and explanations, impact and actionability of feedback and explanations, user understanding of the system, trust in AI, major issues and user concerns, system strengths, user interface, and areas of improvement. These proposed key considerations can help guide the development of effective essay feedback and grading tools that prioritize explainability and transparency to improve usability in computer science education.

A User-Centered Design Approach to Evaluating the Usability of Automated Essay Scoring Systems

Erin E. Hall

(GENERAL AUDIENCE ABSTRACT)

In recent years, rapid advancements in computer science have facilitated the widespread adoption of AI technology across various educational applications, highlighting the need to design and evaluate these technologies with ethical considerations for their impact on students and teachers. Nowadays, there are Automated Essay Scoring (AES) platforms that can instantly provide feedback and grades for student essays. AES platforms are computer programs that use artificial intelligence to automatically assess and score essays written by students. However, not much research has looked into how these platforms work and how understandable they are for users. Specifically, AI explainability refers to the ability of AES platforms to provide clear and coherent explanations of how they arrive at their assessments. Algorithm transparency, on the other hand, refers to the degree to which the inner workings of these AI algorithms are open and understandable to users. To fill this gap, we conducted a qualitative study on an AI-based essay writing and grading platform, aiming to understand the experiences of students and graders. We wanted to explore how clear and transparent the platform's feedback and explanations were. Participants shared their thoughts through surveys, interviews, and a focus group. The study uncovered important factors to consider when evaluating AES systems. These factors include the clarity of the feedback and explanations provided by the platform, the impact and actionality of the feedback, how well users understand the system, their level of trust in AI, the main issues and concerns they have, the strengths of the system, the user interface's effectiveness, and areas that need improvement.

By considering these findings, developers can create better essay feedback and grading tools that are easier to understand and use.

Dedication

Dedicated to my friends and family who have supported me throughout my postgraduate journey.

Acknowledgments

First, I would like to thank my advisor, Mohammed Seyam, whose mentorship and guidance have been invaluable over the past two years. Your insights and feedback have not only refined the quality of my writing, but have also shaped my perspective as a researcher.

I would also like to thank my co-advisor, Dan Dunlap, for the generous amount of time and thought he put into giving me feedback and pushing me to consider different angles and perspectives. Your enthusiasm has been infectious and motivational throughout this process.

Additionally, I would like to thank Daphne Yao, a member of my committee, for her valuable contributions as a professor and during the defense. Your thoughtful insights and constructive feedback greatly enhanced the rigor and quality of my thesis.

Finally, I would like to thank Jessica Tenuta, representing Packback, for giving me the opportunity to work with Deep Dives. Your dedication to advancing educational technology and your willingness to share your expertise have been instrumental in broadening my understanding and knowledge of AI in education.

Contents

List of Figures

List of Tables

List of Abbreviations

AES Automated Essay Scoring

AI Artificial Intelligence

LLM Large Language Models

ML Machine Learning

SUS System Usability Scale

UI User Interface

XAI Explainable AI

Chapter 1

Introduction

1.1 Motivation

In recent years, there have been rapid advancements in computer science including, but not limited to increased capabilities of machine learning models like Large Language Models and the accessibility of large datasets. These advancements have resulted in the widespread adoption of artificial intelligence (AI) technology, such as ChatGPT, across a variety of applications. In the educational sector, as AI has become increasingly adopted, it is imperative to design and evaluate this technology with concern for both ethics and its impact on students and teachers.

In computer science education, technical writing skills and effective communication play a pivotal role in improving students' professionalism. As computer science curricula increasingly incorporate writing assignments, Automated Essay Scoring, and other AI-driven systems have gained significance as valuable tools for grading and providing feedback. These systems offer real-time feedback and grades, enhancing the learning experience for computer science students while helping students to adhere to a set of foundational writing conventions and promoting a higher quality floor.

Despite the growing adoption of AES platforms, limited research has specifically focused on AI explainability and algorithm transparency and their influence on usability within

the context of computer science education. Understanding the impact of AI explainability and algorithm transparency on the usability of AES platforms is crucial for their successful integration into computer science education. Exploring these factors not only enhances transparency and trust but also facilitates informed decision-making for both instructors and students, ultimately improving their learning and teaching experiences.

1.2 Research Questions

In order to address this research gap, our study aims to investigate the characteristics of an effective AI-driven platform and develop a set of key evaluation considerations to assess the usability of such tools. We conducted a literature review to analyze the current state of research in this area and collected data on user experiences and perceptions using Packback Deep Dives, an automated essay grading and feedback platform. The primary focus of this study is to examine the usability aspects related to AI explainability and algorithm transparency and their implications for computer science education. The following research questions guide this study:

1. How do explainability and algorithm transparency techniques affect the overall usability and user experience of an AI-based essay feedback system?

2. How do graders perceive the integration of an AI-based essay feedback system into their grading process, and what are factors influencing their acceptance or resistance of automated feedback?

3. What are the key components that constitute an effective automated essay scoring system, and how can they inform the development and assessment of reliable grading and feedback tools?

1.3 Contributions

This paper presents a number of different contributions to the fields of computer science, artificial intelligence, HCI, and usability. The contributions of this paper can be summarized as follows:

- **Synthesis of AI Explainability and Algorithm Transparency:** This work provides a synthesis of prior work in the field of AI explainability and algorithm transparency and identifies gaps in prior research. This paper also details the complex relationship between AI explainability, algorithm transparency, and usability.
- **Usability Evaluation for Packback Deep Dives:** A usability evaluation was conducted on Packback Deep Dives through a qualitative study in relation to AI explainability and algorithm transparency, and the System Usability Scale.
- **Identifying Usability Challenges in AI-driven Essay Feedback Platforms:** This paper identifies a set of usability problems in the context of AI-driven essay feedback platforms and uncovers the complexities of integrating artificial intelligence and automation into writing-based educational contexts.
- **Considerations for AI-driven Platform Development and Evaluation:** Finally, this paper presents a list of key components to consider when designing and evaluating AI-driven platforms that prioritize AI explainability and algorithm transparency to improve usability.

Chapter 2

Related Work

This chapter outlines areas of prior research that are necessary to address before describing the research methods of this paper. This includes behavioral design, machine learning (ML) feedback, explainable AI, AI literacy, algorithm transparency, AI in education, usability in AI-driven systems, and Packback Deep Dives.

2.1 Behavioral Design

Looking at the relationship between human behavior and technology, previous design principles have focused on the effectiveness of designing technology to meet the needs of the user [1]. While this approach still has its applications, advancements in technology and the way developers approach design has given an alternate design principle an opportunity to manifest itself in many modern day websites and applications. Behavioral or persuasive design explores how design can be used to shift or alter human behavior [2] [3], compelling users to align their actions with the needs of the technology. By leveraging persuasive design techniques, automated essay scoring platforms can encourage students to engage in the writing process, adhere to writing conventions, and produce high-quality essays. However, it is crucial to consider ethical implications and ensure that behavioral design principles are used responsibly, avoiding manipulative practices [4, 5]. By aligning user behavior with the goals of the system, these platforms have the potential to improve writing skills, facilitate efficient

grading, and promote a positive learning experience for both students and instructors.

That being said, behavioral design can be both intentional and unintentional. One way designers often unintentionally alter user behavior is through poor design [6]. Failing to provide the user with sufficient information on the platform they are interacting with can prevent them from making informed decisions, ultimately impacting the usability and quality of their user experience. This is an important consideration in educational-based platforms, as different user groups have different values and interests, including designers, instructors, graders, and students. Understanding these diverse perspectives is vital, as the goal of such platforms is to promote learning and optimize feedback efficiency and effectiveness. Considering the influence of design on user behavior and mitigating any negative impacts of design becomes essential in helping these platforms achieve their goals.

Because algorithms have the capability to influence user behavior, there is a growing need to inform users on how ML is altering their perception of reality [7]. Increasing user's AI literacy, which can be partially achieved through algorithm transparency, can help them to engage with these systems in an informed and effective manner. The following sections will provide more comprehensive overviews of both topics.

2.2 Machine Learning Feedback

Machine learning is a subset of artificial intelligence that has the ability to make predictions based on input data and "learn" over time while increasing in accuracy as new data is fed into the algorithm [8]. Machine learning is a key component of many automated essay scoring systems, enabling the generation of real-time feedback and grades for student essays. Understanding the underlying technical aspects of ML algorithms and feedback generation can help effectively evaluate the usability and effectiveness of these automated platforms.

ML algorithms, such as deep learning models, utilize statistical methods and pattern recognition to make predictions based on input data [8, 9]. These algorithms learn from large datasets and improve their accuracy over time, mimicking human intelligence in decision-making processes [8]. However, the complexity of ML algorithms can result in "black box" models, where it becomes challenging to understand how the algorithm arrives at its conclusions [10].

For example, one specific ML algorithm commonly used in AES systems is BERT (Bidirectional Encoder Representations from Transformers). BERT is a deep learning model that utilizes transformer architectures to understand the contextual relationships within text [11]. BERT-based models, with their multiple layers of hidden units and intricate computations, pose challenges in interpreting their decision-making process due to the high dimensionality of the representations, complex interactions between layers, and non-linear activation functions used in their architecture, limiting their explainability and interpretability [11]. This lack of interpretability poses a challenge to improving the explainability and transparency of AES platforms, and is common across a variety of different ML models.

2.2.1 Black Boxes

More complex ML algorithms can reach a point where even their designers cannot conceptualize how the algorithm is reaching its conclusion. These algorithms are called "black box" models [10]. A common factor that contributes to the complexity of an ML algorithm is the use of large amounts of input data, known as "big data" [12]. This, coupled with an algorithm's ability to find intricate relationships between data points and improve their predictions at extremely fast rates, prohibits humans from being able to easily interpret the algorithm's decision making process [12].

This presents a tradeoff between accuracy and explainability. In applications where feedback is important and accuracy is less critical, complex algorithms like these can still be beneficial. However, in applications where high accuracy is necessary, there is more value in prioritizing explainable, consistent, and transparent models. Using black box algorithms in contexts where there are high stakes, such as in the criminal justice system [13] or in contexts where a student's grade can be influenced, introduce major social and ethical implications that must be considered.

More specifically, in the context of automated essay scoring systems, black boxed models can score essays with a reasonable level of accuracy when compared to previously scored datasets. However, they present issues when it comes to score interpretability. In instances where these models cannot provide justifications for their scoring decisions, users may face difficulties in trusting the system, potentially leading to missed opportunities for learning and growth. These models also raise ethical concerns, including the potential introduction of undetected bias [14].

2.3 Explainable AI

To address the challenge of AI explainability, research is focused on developing techniques for opening black box models and providing insights into their decision-making processes [15, 16]. XAI techniques aim to uncover the inner workings of ML algorithms and provide explanations for their predictions, bridging the gap between the technical complexity of AI models and the users' understanding [15, 17]. In the context of AI-based essay scoring systems, XAI can help students and instructors understand how the system evaluates their writing, fostering improved writing skills and building a deeper understanding of AI. It can also benefit users by giving them the "right to explanation" [18].

That being said, achieving explainability in AES platforms is particularly challenging when ML algorithms, such as BERT, operate as black box models that output highly accurate predictions but pose limitations when it comes to attempting to trace the decision-making process [12]. [15]. The complex internal mechanisms of these models make it difficult to directly interpret their decisions. Researchers are actively investigating methods to open the black box and extract meaningful explanations from such algorithms, ensuring that the feedback provided by AES systems is understandable and useful to users [15].

2.4 AI Literacy

AI literacy involves a general understanding of how artificial intelligence works and what an algorithm is doing. It is defined to be "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" [19]. This oftentimes involves building an understanding of AI outside of specific instances and applications the user may encounter in the real world. The main goal of AI literacy is to help users understand AI, so even if no information is provided on how a specific system's algorithm works, the user has a general understanding of what they are interacting with. This helps bridge the gap between the mathematical principles that drive AI system and the user's mental model of how the system actually works.

AI literacy is a broad concept–this understanding of AI doesn't necessarily have to be specific to any one system, but it can be. It is a combination of a general understanding of AI that can be used to make sense of systems that are not offering explanations and a more specific understanding of how certain systems work.

Beyond helping users collaborate more effectively with AI systems and increasing the us-

ability of the system, AI literacy can also help build trust between a user and a system [19]. Knowing how a system works, where it performs well and poorly, and what its limitations are help users interact with these platforms more confidently and informatively [20].

To improve AI literacy, there is increasing research on what competencies are needed to effectively interact with AI and how technologies themselves can help users understand AI. [19]. Additionally, repeated interactions with a system contribute to AI literacy [21]. Developers can help increase their users' AI literacy by providing users with feedback that indicates how the algorithm is producing its output. This practice is called algorithmic transparency and is outlined in the following section.

2.5 Algorithm Transparency

One approach in harnessing XAI techniques is to provide simple and understandable pieces of feedback that cater to users' limited attention spans and potential lack of interest in technical details [15]. By translating the mathematical concepts and prediction-making techniques of ML into human-like narratives, users can gain insights into the decision-making process of the AES system [15]. This approach is referred to as algorithmic transparency, and its value lies in its ability to empower users to make informed choices and judge the potential consequences of the system's outputs [21]. Like user-centered explainable AI, algorithmic transparency does not necessarily need to be detailed and technical, but it needs to convey just enough about the algorithm to keep the user informed and interested [22].

These explanations can help users evaluate the accuracy of an algorithm, as contextual information can help users understand the rationale behind certain algorithmic decisions. This can help reduce the likelihood of a user jumping to the conclusion that an algorithm is "flawed" or "ineffective". Humans are generally skeptical of AI due to a lack of confidence in

themselves and lack of trust in the system, so algorithmic transparency has the potential to help mitigate these doubts in the system [20]. It should be noted, however, that too much algorithmic transparency can have the opposite effect by uncovering system errors that leave users skeptical about the accuracy of the model. Therefore, it has been suggested that "users may benefit from initially simplified feedback that hides potential system errors and assists users in building working heuristics about system operation" [23].

An example of algorithmic transparency in the context of AES is in the Criterion automated essay scoring model. Criterion scores, evaluates, and offers feedback on essays [24]. One type of feedback offered by Criterion's advisory component suggests that a student's essay might be off-topic. The system offers a level of algorithmic transparency by indicating that the essay might be off topic due to the fact that it "does not resemble other essays written about the topic" [24]. This simple piece of feedback uncovers an essential component of the algorithm's decision making process: that it compares submissions to other essays written about that topic. This can very subtly help build a user's AI literacy by uncovering certain aspects about how an AI system works.

There are two parts to effectively convey the output of an ML algorithm: the feedback and the explanation. In the context of AES, feedback consists of a score or value assigned to the quality of writing. An explanation of this feedback describes how the algorithm determined that score. Both components are essential to helping users effectively interact with AES systems, as feedback without explanations lack the context needed to help users learn from their scores and critically evaluate the output of the algorithm [25].

That being said, achieving algorithmic transparency in AES systems can be challenging, particularly when utilizing complex machine learning algorithms. The challenge of algorithmic transparency lies in the fact that many ML algorithms are black boxed. While Explainable AI attempts to mitigate some of these obstacles, black boxed algorithms present a barrier

to algorithmic transparency that can negatively impact the usability of a system. There are currently various research efforts that are exploring methods to incorporate user-centered explanations, simplifying technical concepts and presenting them in a manner that is accessible to users with varying levels of AI literacy [15, 19].

2.6 AI in Education

When it comes to education, there are many new research efforts on how to automate parts of the learning and grading processes. Explainable AI plays an essential role in the domain of AI in education, as it ensures students have access to a transparent grading system that will help them learn and improve upon receiving feedback. Additionally, instructors can benefit from explainable models as they can help them understand the reasoning behind AI predictions and use this to override any evaluation principles that might not align with their personal teaching styles [26].

2.6.1 The Role of AI in Education

Summative Assessment vs. Formative Feedback

Human evaluation of writing that is particularly concerned with content is a subjective practice. Therefore, even an AI algorithm that could perform with perfect accuracy on the set of principles it was trained on might conflict with alternate teaching and evaluation styles. Humans are also particularly hesitant to fully trust AI generated output due to unfamiliarity and skepticism of accuracy [20]. On top of this, these algorithms can introduce bias based on the demographics and backgrounds of the students they are evaluating [14]. Thus, the practice of using AI for summative assessment is not an approach that is widely adopted, as its value lies in formative assessment and to provide added guidance to assist in learning.

2.6.2 Automated Essay Scoring

Automated Essay Scoring (AES) is a technique used to evaluate and grade written essays [24]. AES systems started out being powered by statistical models, but have evolved to use

natural language processing (NLP) and Bayesian text classification, among others [24, 27]. These advancements have helped offer timely and personalized feedback to students, fostering continuous improvement and enhancing learning outcomes [28].

AES systems have the potential to be of great assistance to teachers as well, as they are able to provide feedback on essays at a rate much faster than humans are capable of. AES, when paired with effective formative feedback that includes sufficient explanations can help students simultaneously grow their writing skills and knowledge of AI [28]. However, AES has been critiqued by its accuracy, potential to introduce demographic-related bias, and the absence of a human in the process [14]. Thus, teachers still present immense value in the classroom, as they can help mitigate some of these drawbacks. Thus, the value of AES lies not in its ability to replace teachers as graders, but to be used as a supplement to help speed up certain processes where there is no added benefit having a human in the process, thus leaving instructors with more time to devote to teaching [28].

2.7 Usability in AI-Driven Systems

The usability of automated essay scoring systems focuses on creating user-friendly interfaces that enhance the effectiveness, efficiency, and user satisfaction in assessing and providing feedback on student essays [29]. While there is no one size fits all definition of usability, as its success depends on the goals of the system and the needs of the users, usability evaluation is a key factor in determining the effectiveness of the users' interaction with the system [30]. The System Usability Scale (SUS) is a widely accepted simple tool for measuring usability [31]. By collecting user feedback and assessing user perceptions through tools like the SUS, developers can pinpoint areas for improvement and tailor their systems to better serve the needs of instructors and students. Despite several advancements in improving and

evaluating the usability of automated essay scoring systems, several challenges persist, such as addressing the need for transparency and explainability in the algorithms that drive these systems.

2.8 Deep Dives

This section provides an overview of Packback Deep Dives, the platform employed in the study, highlighting its features and its role in supporting instructors and students in the grading and learning process.

Deep Dives is designed to facilitate the assessment and feedback process for written assignments and is designed for two primary user groups: students and instructors. While Deep Dives offers features for both groups, this paper is focused on its role as a grading assistant for instructors, and thus is concerned with the feedback provided to instructors to help assist with generating grades. However, to gain insights into how instructors use the platform effectively, it is essential to also consider student perspectives.

For students, Deep Dives provides instant feedback as suggestions on how to improve their essays. This platform offers two primary features for students to help ensure that their essays are submission-ready: a writing assistant and a research assistant. The writing assistant provides feedback on six categories, including grammar & mechanics, wordcount & depth, flow & structure, repetitiveness, research quality, and formatting. The research assistant offers automatic citation generation and provides credibility feedback on all cited sources.

For instructors, Deep Dives automates tedious aspects of the grading process to allow graders to focus on evaluating the content and ideas presented by students. In contrast to other AES tools that focus on generating scores for essays, Deep Dives follows a human-machine

teaming approach where instructors work with a semi-automated tool to produce grades. With Deep Dives, the automated feedback is supposed to be used as a tool to assist with assigning scores, rather than the sole determinant of the final grade. Deep Dives does not replace human graders; rather, it leverages the complementary strengths of both humans and machines to produce more efficient and accurate grading results. Deep Dives provides auto-suggested scores for six different mechanics-related categories and offers customizable rubrics, which graders can accept or override. Additionally, Deep Dives flags certain components of students' writing, including profanity and repetitiveness. Deep Dives also offers Smart Highlighting, which helps indicate where students have addressed the required "guiding questions" from the prompt.

2.8.1 Overview of Capabilities and UI

Instructors

To create a Deep Dives assignment, instructors must first add a prompt with optional details such as guiding questions and recommended resources. The instructor is then responsible for creating a smart rubric, where they can set a total number of points allotted for the assignment. This includes two types of grades: content & ideas, which are entirely manual grades, and requirements, which include the AI-generated grades. Instructors are able to choose assignment open and deadline dates, and can choose to allow late submissions.

When configuring the assignment, there are five automated categories that instructors can choose to include in the grading rubric. These include word count & depth, grammar & mechanics, flow & structure research & citations, and formatting. Instructors can set a minimum and maximum word count, as well as a minimum number of citations. The UI for creating a new Deep Dives assignment can be found in Figure 2.1 below.

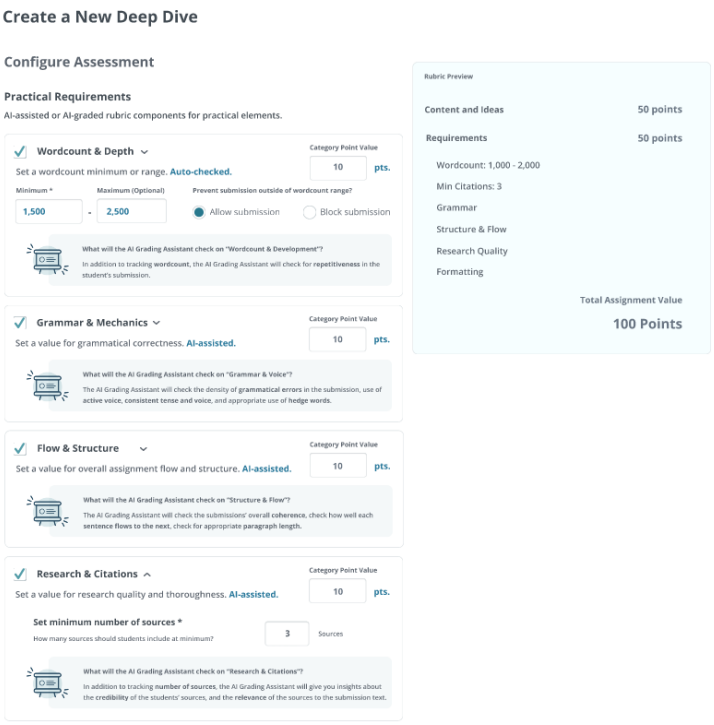

Figure 2.1: Creating a New Deep Dive UI

Currently, Deep Dives does not offer any features for suggesting scores for content and ideas, and the generation of those grades are entirely manual. However, Deep Dives does offer Smart Highlighting, which helps indicate where students might have answered the prompt's "guiding questions". This can help speed the grading process for content by making it easier for an instructor to quickly read and get a high level understanding of the submissions' alignment to key content requirements. Instructors can choose whether or not to include content & ideas in the grades and can assign point values to this section. A four-point scale ranging from "beginning" to "exemplary" is used to grade content & ideas, and instructors can define the criteria for each score. The UI for the content & ideas rubric setup process can be seen in Figure 2.2 below.

Once essays have been submitted, instructors are able to review the automated scores and submit their final grades. Deep Dives displays whether or not the student has met the word

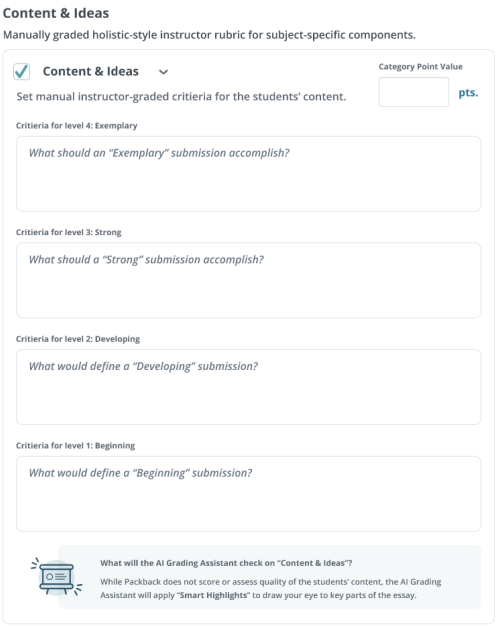

Figure 2.2: Content & Ideas Rubric Setup

count, minimum number of citations, and due date. It then displays the suggested scores, which instructors can edit if necessary. For each section, there is a "see why" button which provides more detailed explanations of automated grades. Lastly, instructors are able to select a score ranging from 1 to 4 for the ideas and content portion of the grade and assign a corresponding point value. The instructor assessment grading view can be seen in Figure 2.3 below.

Students

Deep Dives provides students with a comprehensive writing experience where they can outline, compose, and receive feedback on their essays. The essay writing view includes three tabs: research notes, drafts, and works cited. The research notes tab offers a text editor for outlining and note-taking. The draft tab serves as the main writing area for students to compose their essays, while the works cited tab generates citations for sources added by

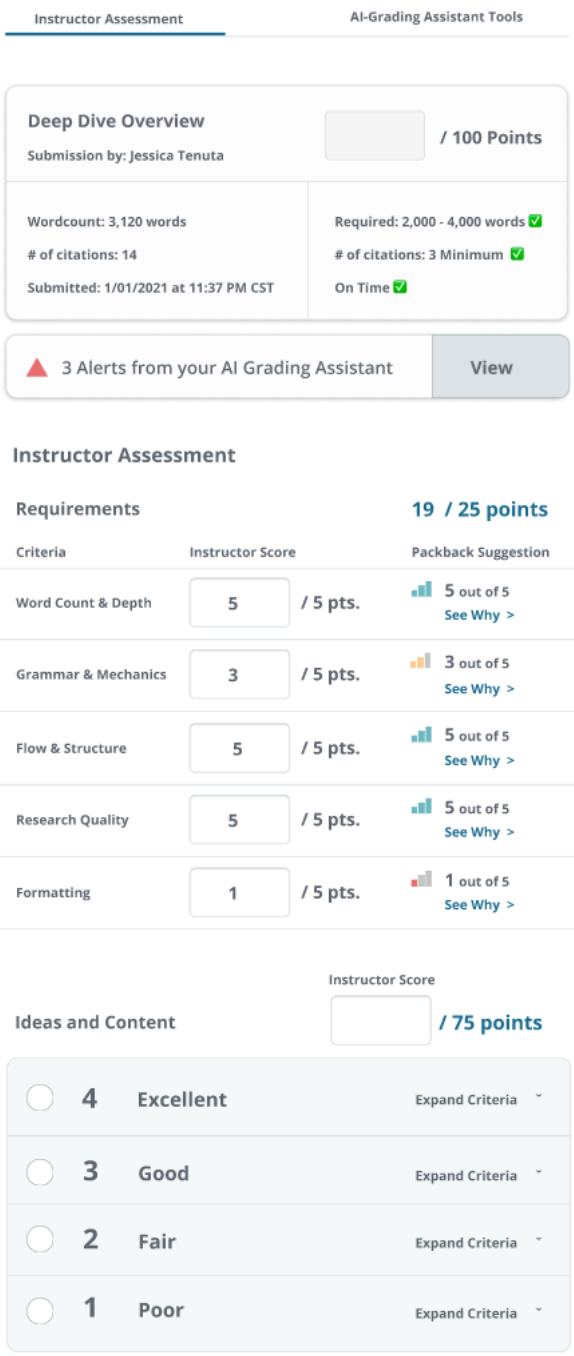

Figure 2.3: Instructor Assessment View

the students. In addition, the platform provides a research assistant and a writing assistant, both accessible from the right side of the essay writing view. This UI can be seen in Figure 2.4 below.

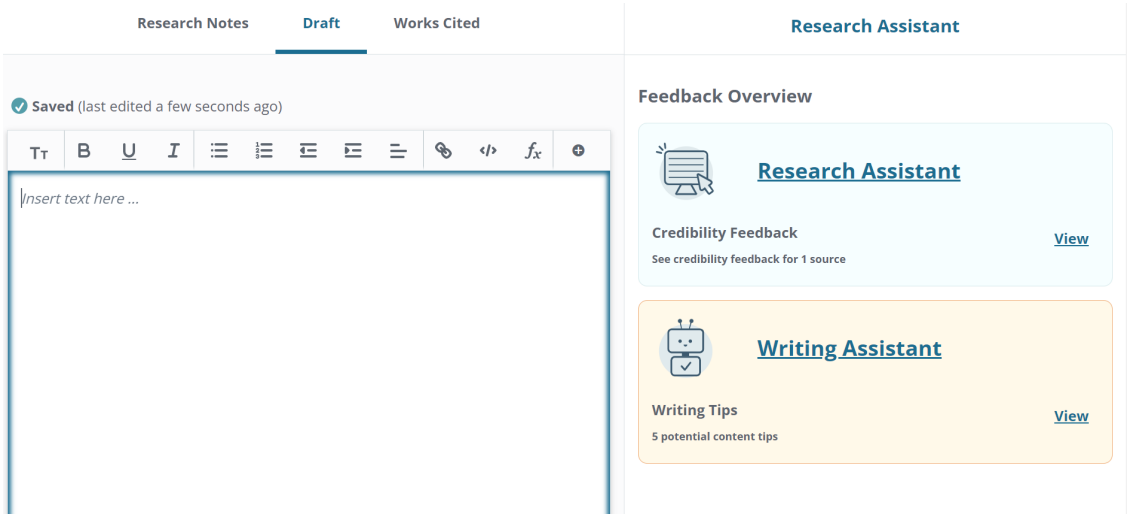

Figure 2.4: Essay Writing View

The research assistant lists all sources added by the student in the works cited tab. It gives each source a credibility ranking and offers the option to include or exclude the citation in references. To get a more detailed description of the credibility of the source, students can click on the credibility ranking. Deep Dives also alerts students if any necessary components of their citations are missing, such as author or publication date. Figures 2.5 and 2.6 provide examples of the credibility check screen and citation alerts, respectively.

The writing assistant tool provides feedback on areas where the essay has the most room for improvement, followed by an option to view all current suggestions. Beneath this, students can see the five automated rubric categories, each with alerts, suggestions, and praises. Under each rubric category, students can see a list of errors, with a button to review and fix those errors. The tool also provides descriptions on how to fix issues, as well as praises for well-done areas. While each rubric category shows only two pieces of feedback at a time,

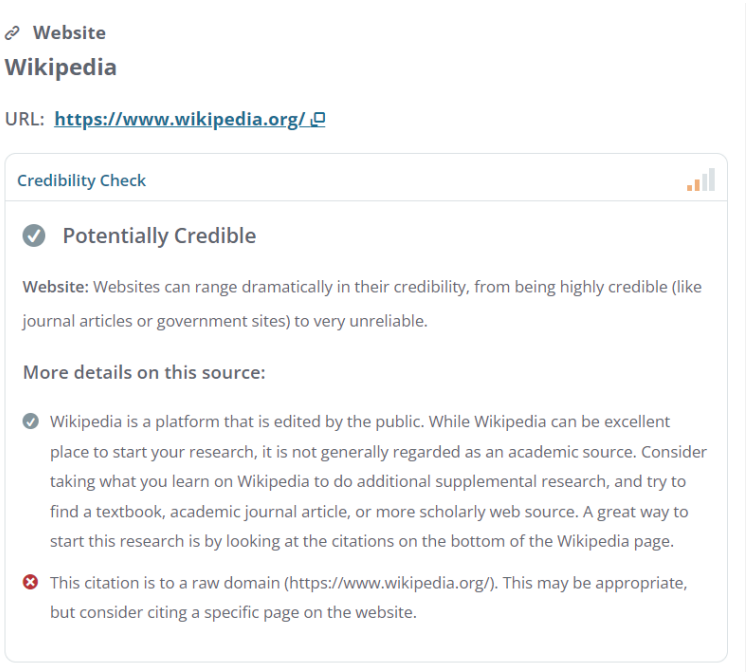

Figure 2.5: Credibility Check

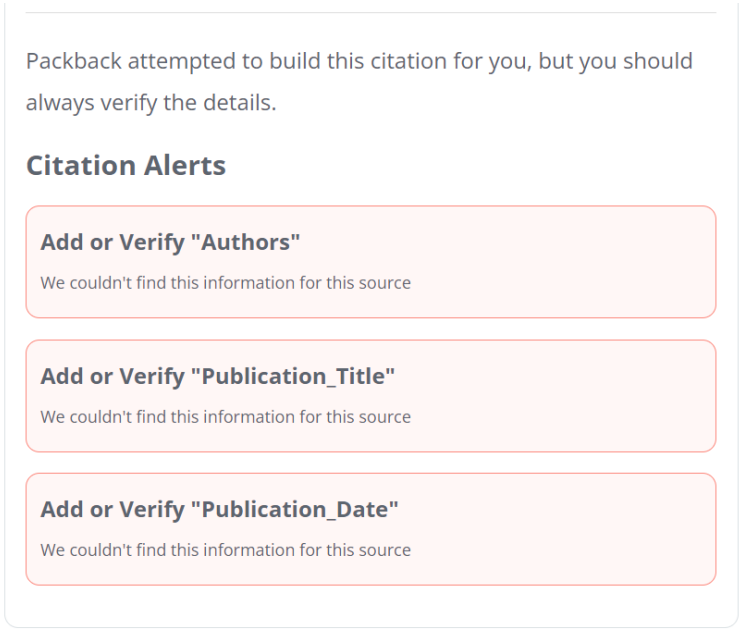

Figure 2.6: Citation Alerts

students can access all pieces of feedback by clicking a button. Although students cannot see exact point values for their grades, they can track their progress with a progress bar that shows if they are close to satisfying all the requirements of a section. An example of feedback given in this section can be seen in Figures 2.7, 2.8, and 2.9 below.

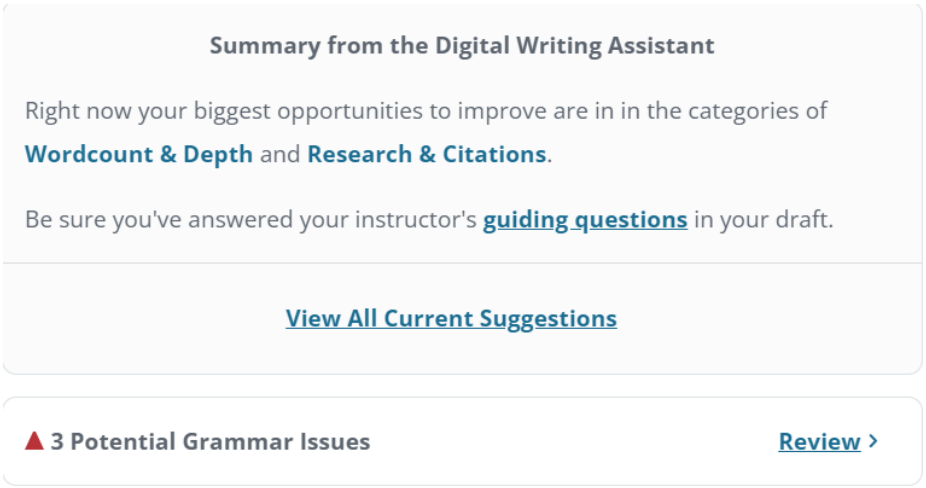

Figure 2.7: Writing Assistant

2.8.2 Automated Scoring Methodology

In the context of Packback's automated essay scoring system, the calculation of the 5 rubric categories, "Grammar & Mechanics," "Word Count & Depth," "Flow & Structure," "Research & Citations," and "Formatting & Presentation", reflects a unique blend of different types of algorithms. Unlike traditional automated essay grading systems, Packback's approach prioritizes real-time, actionable feedback over a singular overall score.

Deep Dive's architecture is primarily programmatic, where the rubrics, structure of how feedback is presented, and highlighting is all "hard-coded". It leverages a variety of different models to deliver each type of feedback, including ML-based, LLM-based, and programmatic Python algorithms. This blend of techniques is strategic, serving the dual purpose of

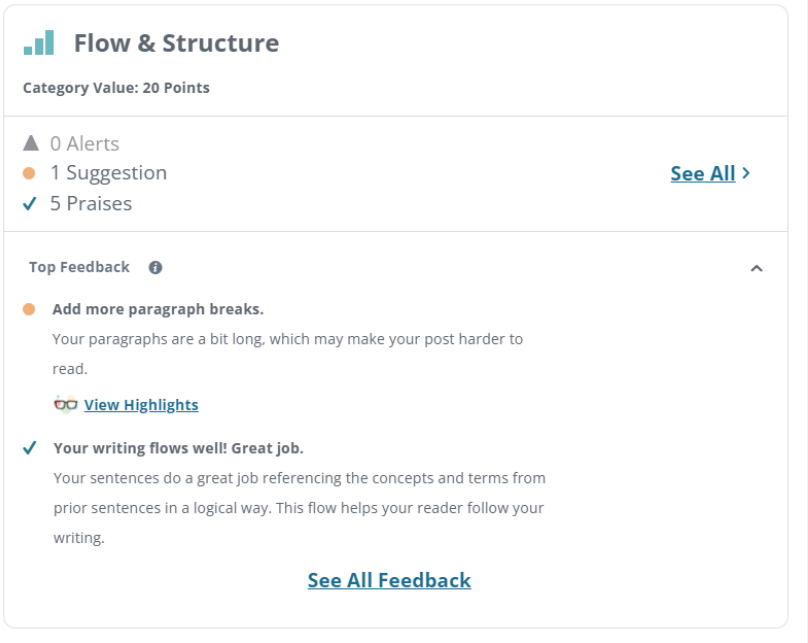

Figure 2.8: Flow and Structure Feedback Overview

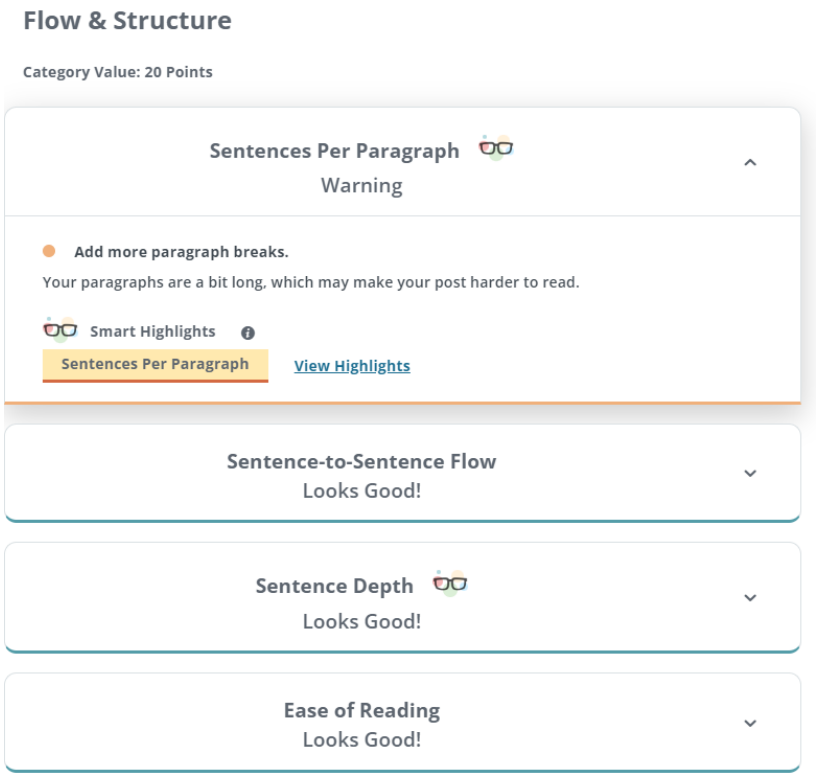

Figure 2.9: Flow and Structure Feedback "See All"

maintaining explainability and mitigating the potential for bias that is inherent in black box LLMs.

For instance, the "Flow & Structure" category score uses an ML-based model adapted from an open-source source model for "Local Coherence." The "Grammar & Mechanics" feedback, on the other hand, employs a combination of programmatic and rule-based systems in addition to an open-sourced ML model. A key decision in this category was to prioritize rule-based feedback over ML-based feedback, a choice driven by the desire to provide users with more explainable feedback that offers comprehensive context.

A key feature of Deep Dive's system is its structured feedback presentation, which enables precise scores to be associated with individual rubric components and the ability to set different automated procedures for assigning these scores.

2.8.3 Addressing ChatGPT in Deep Dives

In contrast to BERT, which is typically used for text interpretation, generative text LLMs raise concerns in the context of essay writing, as they can potentially lead to originality and authenticity issues in student essays. In November 2022, ChatGPT was released, quickly amplifying these concerns.

ChatGPT is an application that leverages GPT3's advanced generative language model to create a chatbot, allowing human users to interact with the model by asking it questions. ChatGPT can summarize, rewrite text, search, cluster, classify, and generate text. It can produce essays that sound convincing and human-like, in a matter of seconds and in various lengths [32].

That being said, when it comes to essay generation, ChatGPT has some limitations. The model may misrepresent specific details or present statements as facts when they are in-

accurate. Additionally, feeding the same prompts to ChatGPT leads to identical essays, as the model is trained on existing content found on the internet, thus limiting its ability to generate original and novel content. ChatGPT may also include inaccurate or made-up citations that weaken the credibility of the essays it generates.

There is some concern that if enough information is given to ChatGPT, it can generate essays that can get full marks on automated portions of Deep Dives. On top of this, these essays would go undetected by plagiarism checkers since the text is being generated and not copied. Currently, there are many efforts that are attempting to leverage these limitations to help detect AI-generated text.

Deep Dives specifically has come out with automated essay writing detection techniques that were released into the instructor experience in January 2023. "Teach with GPT" utilizes AI to detect AI-generated content [33]. ChatGPT was made publicly available towards the end of the semester during which this study was conducted, so there was no automated essay detection in place at that time. It is important to note that these features were not present when this study was conducted.

Despite the limitations of AI-generated essays, there are still identifiable characteristics that humans can detect. Human reviewers are therefore essential in detecting signs of AI-generated content, ensuring that essays meet academic integrity standards, even when they pass automated checks. Collaboratively, both human reviewers and automated systems can effectively combat the threat posed by AI-generated student essays.

To address these issues, Packback has also developed an AI ethics policy that emphasizes the well-being of students and the value of human-machine teaming [34]. The policy underscores the importance of transparency, explainability, and accountability, ensuring that AI decisions can be understood and challenged by humans, and it prioritizes the role of educators in

complementing AI's capabilities rather than seeking to replace them. Through these ethical commitments, Packback works to harness AI's potential while maintaining responsible and effective educational technology. The policy also places a strong emphasis on minimizing bias in AI feedback and actively making efforts to prevent harm.

Chapter 3

Methods

3.1 Approach

This section outlines the methodology employed to address the three driving research questions presented in this paper. A study was conducted in the Fall of 2022 at a large public university in the United States. Several courses at this university utilized an AI-based essay feedback and grading platform to assist with their writing assignments. All participants were part of a course that used this platform, and data on their experiences and perceptions of using this platform were collected during a three phased study over the course of the Fall 2022 semester.

The first phase involved gathering background data on the participants and their relevant experiences through a survey. The second phase aimed to collect the participants' initial impressions of using the AI grading tool, also through a survey. The third phase, conducted after the semester's conclusion, aimed to obtain a more in depth understanding of the participants' experiences using the tool through a survey, individual Zoom interviews, and a focus group. The surveys were intended to be brief and straightforward, allowing for the collection of quantitative data, while the bulk of the qualitative data was gathered from the interviews. The three-phased approach was designed to provide a comprehensive evaluation of the effectiveness of the platform throughout a single semester-long course, and to gather valuable insights into the participants' experiences while they were using this tool.

3.2 Participants and Courses

Participants were recruited to provide a broad range of perspectives across different roles within the academic context. Participants in this study consisted of nine individuals, including five teaching assistants (TA's) for a CS course, one instructor of a non-CS course, and three CS students, all who were currently using the essay grading platform at the time of the study.

The CS course utilized in the study was aimed to help students explore the social, ethical, and professional side of computing. This course required in-depth essay writing as part of the curriculum to help students develop their critical thinking skills and ability to effectively articulate their ideas.

To ensure a diverse range of perspectives, the study also included one participant from a humanities course where essay writing played a key role in the curriculum. While the context of this course deviates from the main scope of the study, their inclusion offers valuable insights into the tool's application in a non-CS academic setting.

All nine participants took part in the individual interviews, but only the six graders completed the surveys, as they were geared towards the instructor-view of the tool. All participants consented to take part in the study, and no compensation was provided for their participation. Participants were asked for consent to have their interviews recorded and transcribed for analysis.

3.2.1 Objectives

The objectives and lists of sample questions for each phase of the study can be seen in Tables 3.1 and 3.2 below:

Table 3.1: Objectives and Sample Questions (Part 1)

Phase	Objectives	Sample Questions
Prelim. Survey	1. To obtain information on participants' prior knowledge and experience with grading, automated systems, and AI. 2. To understand user expectations and gauge their needs. 3. To establish a baseline by exploring participants' past manual grading experiences.	1. Have you graded writing assignments before? 2. Rate your knowledge of AI and ML algorithms. 3. What are your expected advantages of this tool? Disadvantages?
Midpoint Survey	1. To capture participants' initial impressions of the platform. 2. To assess grading efficiency and speed compared to prior experiences. 3. To evaluate participants' trust in automated scores. 4. To gather insights into participants' mental models of the grading algorithm.	1. Describe your initial impressions of the platform. 2. How much do you trust the automated grades? 3. Briefly explain how you understand the grades are generated.
Final Survey	1. To obtain quantitative data on platform usability using the System Usability Scale [31]. 2. To investigate changes in responses between traditional grading and using the tool.	1. How confident are you in grading consistently and fairly throughout the course? 2. What is the average time it takes you to grade a single writing assignment? 3. Did the automated grader grade higher, lower, or the same as you would have?

Table 3.2: Objectives and Sample Questions (Part 2)

Phase	Objectives	Sample Questions
Grader Interviews	1. To clarify responses from previous surveys, including background and specific experiences. 2. To address the system's usability. 3. To understand participants' mental models of the grading algorithm. 4. To assess participants' trust in automated scores.	1. How did using the AI grading platform change your grading process? 2. How well did you understand how certain grades were determined? 3. Explain how you think scores were generated for each rubric category.
Student Interviews	1. To explore students' experiences with writing essays using the platform. 2. To understand their mental models of the grading algorithm. 3. To gain insights into system generated feedback and its impact on their writing process.	1. What was your overall experience with the platform? 2. How well did you understand how certain grades were determined? 3. Can you provide examples of feedback that prompted changes in your writing?
Focus Group	1. To reveal additional experiences and perceptions. 2. To validate the findings from other evaluation methods. 3. To compare mental models of different user types.	1. Was the automated feedback clear to you? 2. Did you understand how the automated system graded you? Why or why not? 3. Can the graders describe how the system works? Do students agree?

3.3 Preliminary Survey

The preliminary survey was sent out prior to the first writing assignment of the semester and prior to any engagement with the platform. Each participant was required to answer a minimum of 10 questions. Additionally, there were 4 optional questions only presented to users with experience grading writing assignments, and 5 other questions presented exclusively to participants with prior experience using automated grading systems. The participants were told that the survey would take around 10-15 minutes to complete.

The survey questions can be found in A.1

3.4 Midpoint Survey

The second survey was sent out after each participant had a chance to become acclimated with the platform, after about half of the semester's writing assignments had been turned in and graded. This survey was used mainly to gauge initial impressions of using the platform and to obtain some preliminary data that would be addressed more in depth in the following interviews. There were 16 questions, and participants were told the survey would take around 10-15 minutes to complete.

The survey questions can be found in A.2

3.5 Final Survey

The final survey was sent out after the conclusion of the semester, but before the focus group and individual interviews with each participant. The first 10 questions of the survey were part of the System Usability Scale [31], while the following questions included questions

repeated from the preliminary survey to examine changes in participants' responses when describing traditional grading vs. grading with the automated tool. The survey had 17 questions and participants were told it would take around 10 minutes to complete.

The survey questions can be found in A.3

3.6 Individual Interviews with TAs and Instructors

Following the end of the semester and completion of the final survey, hour-long semi-structured interviews were conducted with the six participants that used the instructor version of the platform. These interviews took place over Zoom and lasted about an hour.

The interview was conducted more like a conversation than by following a structured guide, however, a list of interview questions was used as a guide to lead the conversation to cover certain topics.

The interview questions can be found in A.4

3.7 Individual Interviews with Students

Similarly, after the conclusion of the semester, semi-structured interviews were conducted with the three participants that used the student version of the tool. These interviews took place over Zoom and lasted about 45 minutes.

These interviews were also conducted more like a conversation to encourage the participants to openly express their opinions and experiences with using the AI-based grading tool. A list of interview questions was used as a guide to lead the conversation towards specific topics. The main goal of these interviews was to get an understanding of student impressions of how

the tool works in order to compare their experiences and perceptions of the tool with those of the TA's and Instructors.

The interview questions can be found in A.5

3.8 Focus Group

The final methodology employed to collect quantitative data after the conclusion of the semester was a focus group. The focus group consisted of three participants from the previous interviews: one student, one Instructor, and one TA. The participants were chosen based on their varying backgrounds and different uses of the platform. The focus group was conducted over Zoom and lasted about an hour.

A discussion guide with a list of questions was used to guide the conversation of the focus group, but participants were still encouraged to converse with each other and share their experiences and thoughts freely.

The focus group questions can be found in A.6

3.9 Data Analysis and Evaluation

The background data obtained in the preliminary survey was collected in order to understand how each participant's past experiences affected how they use Deep Dives and contributed to any bias going into the study.

The system usability scale was used as the primary way to obtain quantitative data on the usability of the system. The System Usability Scale (SUS) is a survey designed to measure the usability of a system throughout a variety of contexts [35]. The survey consists

of ten questions that alternate between statements of positive and negative connotations. Participants were asked to rank each statement on a five point scale ranging from “Strongly Disagree” (0), to“Strongly Agree” (4). The values were then scaled to produce a score out of 100. A score “above a 68 would be considered above average and anything below 68 is below average” [31].

The audio recordings and transcripts of the individual interviews and focus group were evaluated using reflexive thematic analysis conducted by a single researcher [36, 37]. The raw data from interview transcripts were thoroughly reviewed and familiarized to develop a comprehensive understanding of the content and context. An inductive approach was employed to create the codes, allowing for the emergence of patterns and concepts directly from the data itself. The researcher conducted a line-by-line analysis of the transcripts and extracted significant quotes from participants. These quotes were then assigned unique labels as codes that consisted of descriptive keywords or short phrases that encapsulated the key concepts and ideas expressed in the data.

To ensure consistency and rigor in the coding process, a systematic and iterative approach was followed. The researcher repeated this process three times, constantly comparing and refining the codes as new insights emerged. Codes were iteratively reviewed, condensed, reworded, and refined. Finally, each of these codes were organized into broader themes.

Chapter 4

Results

4.1 Participants' Backgrounds and Expectations

The first survey asked several questions to gain a better understanding of the background of the participants in this study. It was administered exclusively to the six instructors and TAs involved in the study.

The survey found that five participants had experience grading writing assignments, some of which were for the same course in which they were currently using Deep Dives. The participants exhibited a range of confidence levels, varying from neutral to very confident, in their ability to grade consistently and fairly, as well as to grade consistently with other graders. The participants had varied responses for which sections of the rubric they prioritized while scoring papers, however, several stated that they focused most on content and original thought. Five of the participants had recognized that they had interacted with AI systems before, although their levels of knowledge of artificial intelligence and machine learning algorithms varied. One respondent had used an automated grading tool, specifically SAGrader, in the past. Four of the respondents were familiar with Packback prior to this study, having used Packback Questions before.

Participants had different expectations for what the advantages of using Deep Dives in essay grading would be. Some stated that they felt it would cut back on time checking grammar,

while others expected it to raise the overall writing quality of essays turned in by students.

In addition to their expectations, participants expressed concerns about utilizing an AI system for grading. Several of the concerns mentioned in the survey are presented below:

> *I believe that style and artistry of writing contribute greatly to the quality of an essay; I am unsure how well an automated system can evaluate style choices and worry some suggestions it returns may even discourage intentional stylistic choices on the students part.*

> *Students might try to "hack" the system.*

> *I think one worry that I have with Deep Dives is its ability to be consistent between students.*

4.2 Themes

Following the completion of the study, 174 initial codes were extracted from the survey responses, interview, and focus group transcripts. These codes were then analyzed for similarities and condensed into a smaller list. 8 broader themes emerged that captured the main findings of the data. Each coded piece of data was organized into one of these 8 themes. The themes are outlined in Table 4.1 below and the number of codes that contributed to each theme are listed in Table 4.2 below. The full list of codes for each theme can be found in Appendix B.

Table 4.1: Overview of Themes

Theme	**Description**
Clarity of Feedback and Explanations	Both instructors and students highlighted the importance of clearer expectations, informative explanations, and specific examples to improve the system's clarity.
Impact and Actionability of Feedback and Explanations	Instructors valued time-saving benefits, but double-checked automated grades due to occasional errors, while students emphasized the importance of clearer and more actionable feedback to prompt effective changes in their writing practices.
Understanding of System	Graders had varying levels of understanding of how the grading algorithm worked.
Trust in AI	Graders had varying levels of trust in the grading algorithm and their trust changed over time.
Issues and Concerns	Instructors questioned the fairness of enforcing a specific writing style, while students expressed frustration with contradictory feedback and desire for more comprehensive guidance and engagement with the AI.
Strengths	Instructors and students both appreciated the automated writing assessment system's efficiency, assistance in refining writing skills, and provision of helpful feedback.
User Interface	Graders found the UI easy to use.
Areas of Improvement	Graders had suggestions for areas of improvement in terms of feedback and added UI features.

Table 4.2: Number of Codes that Contributed to Each Theme

Theme	**Instructor Codes**	**Student Codes**
Clarity of Feedback and Explanations	15	23
Impact and Actionability of Feedback and Explanations	5	12
Understanding of System	19	16
Trust in AI	27	5
Issues and Concerns	9	14
Strengths	6	4
User Interface	9	4
Areas of Improvement	4	2
Totals	94	80

4.2.1 Clarity of Feedback and Explanations

Instructors expressed concerns about the algorithm's grading decisions and inadequate explanations, and emphasized the need for better understanding of expectations. They called for more specific examples and clear definitions to improve feedback clarity. Students, on the other hand, complained about vague feedback, confusing suggestions, and a lack of specific guidance in addressing writing issues. They desired more informative explanations and targeted feedback. Both groups emphasized the importance of enhancing the system's clarity by providing specific examples, clearer expectations, and more informative explanations.

The following quotes from participants highlight some of these concerns:

> *If I had to associate a word with Deep Dives overall, it's probably vagueness. In this situation more information is almost certainly better.*

> *The feedback is pretty broad. It's not necessarily a fine-grained, detailed feedback.*

Additionally, below is a sample of codes from reflexive thematic analysis that contributed to this theme, categorized based on whether they were stated by a grader or a student.

- Grader Codes:
  - Sometimes students got full credit/didn't and it wasn't clear why.
  - Wishes there was more feedback and that they could see specific examples of what was wrong like the students could.
  - Didn't always know what the algorithm was looking for.
- Student Codes:

- Sometimes feedback was contradictory.
- Broad and vague feedback about your writing as a whole was confusing and hard to fix because it wasn't specific enough.
- Word count, number of sources, grammar issues, and credibility were all clear pieces of feedback.

4.2.2 Impact and Actionability of Feedback and Explanations

Instructors acknowledged the system's ability to alleviate tedious tasks such as word count and plagiarism checks, thus helping reduce the amount of time they spent checking those categories. Many of them noted that after encountering errors in the grading logic, they felt the need to always double check automated grades to ensure accuracy. Students reported varied responses to feedback, with some disregarding the feedback when they were unable to understand how to address it effectively. While certain pieces of feedback prompted immediate changes in their writing, feedback requiring more effort was often left unaddressed. Both instructors and students identified instances where the feedback seemed unreasonable or confusing, leading to its dismissal. Findings suggest that improving the actionability and clarity of feedback would enhance its impact on students' writing practices.

The following quotes from participants highlight some of these concerns:

> *If the scores were aligned with what I expected, I would be fine with it, but if the scores were off to me, then I would go back and double check it.*

> *When I thought the feedback mattered, I did act on it.*

Additionally, below is a sample of codes from reflexive thematic analysis that contributed to

this theme.

- Grader Codes:
  - Would adjust unclear points.
  - Double-checked first few times and found errors, so felt they always needed to double-check.
  - Completely stopped looking at reference grades because they were always wrong.
- Student Codes:
  - Would ignore AI feedback when they couldn't figure out how to fix it.
  - Low effort changes where feedback was specific would make them change their writing.
  - Sometimes feedback was difficult to act on.

4.2.3 Understanding of System

Instructors highlighted the influence of their prior experience with similar tools in comprehending the system's functionality. They recognized the system's capability to analyze grammar and mechanics using Natural Language Processing (NLP), as well as the presence of specific grammar and structure rules. However, several graders had misconceptions related to students' visibility of concrete scores and what type of feedback students were receiving. While instructors sought a deeper understanding of the algorithm to enhance their grading and ability to expand on feedback, there were still uncertainties about how different rubric categories, specifically formatting & presentation, were calculating grades. Students, meanwhile, relied on trial and error, discussions with peers and professors, and guesswork to

decipher how the system operated. They made assumptions about the algorithm based on feedback and patterns, noting certain qualities of sources that were deemed highly credible and that the grammar algorithm followed a set of rules. Misconceptions regarding how the flow & structure category and credibility assessments were calculated were also apparent. Students expressed a desire for a clearer understanding of the algorithm to aid their writing process and navigate grading outcomes more effectively. Below is a quote from a participant that highlights these concerns.

> *If I had a deeper understanding of exactly how the algorithm works for each subcategory, I could sort of click around each paper, figure out which subcategories I trusted and which ones I don't, and what to look for for each one.*

Additionally, below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - Previous experience with similar tools helped them figure out how the system worked.
  - Having a deeper understanding would help them determine which categories they trusted more.
  - Understood how the system worked based on a little information Packback gave, reading, and picking up patterns.
- Student Codes:
  - Knew how algorithm worked from guesses, trial and error, and discussing with professors and other students.

- Noticed that scores would naturally improve as they wrote more.
- No idea how flow & structure was calculated, played around with it a lot and best guess is it has to do with the amount of words in a sentence.

Lastly, Table 4.3 provides an overview of the rubric categories and paraphrased quotes from participants describing their perceptions of how scores were calculated. It is worth noting that there were diverse mental models among participants regarding the functioning of the system.

4.2.4 Trust in AI

Instructors expressed varying degrees of trust in the system, which increased over time as they became more familiar with its workings. They acknowledged the importance of double-checking scores for fairness and exhibited greater trust in the system's grammar and mechanics evaluation compared to its assessment of flow and structure. While some instructors had disagreements with the system's grading decisions, they generally recognized its consistency with their own evaluation approach. Students, on the other hand, tended to trust the AI without questioning it extensively. They often received higher grades than the tool initially gave them, learning to understand not to rely solely on the automated system's scores, as graders have the ability to override. Overall, different perceptions of fairness and trustworthiness were associated with different individual grading categories and specific grades. Instructors occasionally overrode grades, whereas students believed the AI was generally fair. These findings highlight the complexity of trust dynamics between instructors, students, and the automated writing assessment system, emphasizing the importance of transparency and a clear understanding of the system's capabilities to foster trust. The following quotes from participants highlight these concerns:

Table 4.3: Perceptions of how the Algorithm Works

Rubric Category	**Participant Mental Models**
Wordcount and Depth	• Score is half word count, half length of paragraphs. • Word count looks at white spaces to determine words and counts "tokens". Not sure what depth was doing. • Gave higher scores to five sizeable paragraphs, opposed to smaller paragraphs. • Looked for spaces or groups of words and how many words were in the text. • Students lost points if they repeated themselves. • Packback would take off points if they went way over the word limit. • Scores were calculated by number of words divided by wordcount, times the number of points association to that section.
Grammar and Mechanics	• Scores were related to Kaplan scores and also used something like Grammarly with set rules. • Used Natural Language Processing (NLP) by feeding a dataset into the algorithm so it could learn words and grammar given rules. • Graded based on what it was trained on; commonalities between good papers and grammar rules. • Wasn't sure, thinks there were certain subcategories.
Flow and Structure	• Flow focused on transitional phrases. Wasn't sure about structure. • Structure looks for 5 paragraph essays. Flow looks for how well sentences lead into each other and if they have transitional words. • Looked for concrete details like how many sentences per paragraph or the presense of transition words. • Looks at how long paragraphs are, flow between topic sentences, and connections between one paragraph to the next. • Looks at words per sentence and words per paragraph.
Research and Citations	• Deducted points if citation was missing date. Quality was calculated into grade and in text citations didn't count. • Deducted points if certain things were missing in citations. Had hard-coded rules that determined grades • Flags citations if they are not reliable sources • If source was missing a title or author, points were deducted. • Determined if a source was credible by a hard coded list. • More sources did not improve grades if the credibility wasn't high.
Formatting and Presentation	• Not sure how this was calculated, never saw low scores. • Thinks that instructors could have presets for how they wanted papers formatted. • Similar to word count and depth where it pulls concrete details out of the text. • Student essays that contained a header with the course and instructor received better scores.

My trust grew over time. Since this was my first time using it, I was trying to read the papers carefully and evaluate what I could trust it on and what I couldn't trust it on. So building trust took time.

In my mind I think of writing as such a human activity and communication is so nuanced, so I definitely had that bias going in of not being sure how well this is going to capture those elements.

When it was introduced we were told you can fully trust it, it was a trial of using it, but it was also described as something really great and really accurate. I think if I knew then what I know now, I would have been more cautious with using it.

I just kind of trusted whatever the system was doing. I generally don't spend a whole lot of time taking points away or providing too much feedback in terms of writing just basic writing stuff.

I didn't inherently trust it, but I also didn't inherently think it was wrong. I felt that it needed to be checked, but not necessarily that it was giving back the wrong scores.

Additionally, below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - Got more comfortable with the algorithm as time went on.

  - Trusted grammar & mechanics scores the most.
  - Didn't trust it less over time but assigned it a different role–to assist grading and not replace it.

- Student Codes:
  - Usually didn't question the AI.
  - Learned not to take Deep Dives scores at face value.
  - Usually got better grades than Deep Dives made it seem like.

4.2.5 Issues and Concerns

Instructors questioned the fairness of forcing students into a specific writing style, particularly in relation to the subjective nature of evaluating flow and structure. They occasionally disagreed with the system's assessment and faced dilemmas regarding whether to override scores. Students expressed frustration with contradictory feedback, a desire for a comprehensive view of feedback, and a wish for deeper engagement with the AI. They often found the guidance provided by the system to be restrictive and limiting to their own writing process. Additionally, concerns were raised about the system's assessment of source credibility, exceptions to grammar rules, and contradictory feedback on paragraph length. These findings underscore the need for improved clarity, consistency, and flexibility within the automated writing assessment system to address the identified issues and alleviate concerns of both instructors and students.

The following quote from a participant describes one such concern:

It kind of seems like it's kind of trying to pigeonhole different styles of writing into a more straightforward and basic style of writing, which sometimes is a good

thing, and sometimes it's a bad thing.

Below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - Questions if it is fair to put students in a box and force them to write in a certain way.
  - Flow & structure is pretty subjective.
  - More boring papers were more likely to get full credit.
- Student Codes:
  - Deep Dives weren't necessarily difficult but they were extraneous in that the way that the software guided their writing prohibited her own writing process and her wanting to expand on certain thoughts.
  - Wishes they could see all pieces of feedback at once so they could fix the easier ones first.
  - Didn't like how it essentially made you dumb down your writing.

4.2.6 Strengths

Instructors acknowledged the system's ability to expedite the grading process, providing a second opinion that could potentially save effort. They appreciated features that checked for repetitiveness and toned down flowery writing that bordered on being unreadable. Students also recognized the system's strengths, including the small checks that aided them in refining their writing. They found value in the word count objective and the visibility of required

citations. Additionally, they appreciated the dynamic feedback provided during the writing process. These findings highlight the strengths of the automated writing assessment system in terms of efficiency, assistance in improving writing skills, and the provision of useful feedback. Understanding and leveraging these strengths can contribute to the effective use of the system in supporting student learning and enhancing the writing assessment experience.

The following quote from a grader highlights one example of these strengths:

> *I kind of second guess myself a lot when I'm grading so just having that reassurance of the Packback score agreeing with me was nice.*

Below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - Made grading faster.
  - Liked having a second opinion.
  - Sometimes it was nice to have AI tone down flowery writing that was borderline unreadable.
- Student Codes:
  - Likes the small checks Deep Dives has.
  - Helps lower kids more–students who already write well will continue to write well.
  - Likes the dynamic feedback while writing.

4.2.7 User Interface

Instructors generally found the UI to be easy to use and navigate, with feedback readily accessible. However, they noted that the feedback could be somewhat hidden and required multiple clicks to access the flagged elements. Some instructors encountered challenges in finding explanations and found it time consuming that they had to manually transfer grades to the gradebook. Conversely, students reported finding everything within the UI to be easily accessible. They often left their writing unchanged due to difficulties in addressing confusing feedback errors. While the UI was generally user-friendly, students noted occasional issues such as small font size and difficulties in viewing more than two pieces of feedback. These findings shed light on the strengths and limitations of the system's user interface, emphasizing the importance of clear and intuitive design to enhance the user experience for both instructors and students.

The following lists a few bugs and issues that were brought up in interviews:

- Sometimes the page would resize and be unresponsive for a second.
- If you enter a grade into a field and then scroll without exiting the field, it changes the number.

Below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - UI was easy to use.
  - Sometimes couldn't find the explanations.
  - Took 3 or 4 clicks to see what was being flagged.

- Student Codes:
  - Everything was easy to find within the UI.
  - Couldn't figure out how to see more than two pieces of feedback.
  - Usually left their writing as is because they couldn't't figure out how to get the confusing feedback errors to go away.

4.2.8 Areas of Improvement

Instructors expressed the desire for inline comments, allowing them to provide more specific and contextualized feedback. They also raised concerns about the system flagging grammar issues that were not actual errors and called for an enhanced plagiarism check. Additionally, instructors wished for clearer visibility into the system's prompts and instructions provided to students. Students, on the other hand, expressed the need to view all feedback at once to prioritize easier revisions. They also wished for deeper engagement with the system, with a desire for more informative feedback. These findings indicate the need to address these areas of improvement to enhance the functionality and effectiveness of the automated writing assessment system.

Below is a sample of codes from reflexive thematic analysis that contributed to this theme.

- Grader Codes:
  - Wants inline comments.
  - Wants a better plagiarism check.
  - Wishes they could see what the AI was asking the students to do.
- Student Codes:

- Wishes they could see all pieces of feedback at once so they could fix the easier ones first.
- Wishes Deep Dives gave more information because they can't have a discussion with AI like she could with a professor.

4.3 Usability of Deep Dives

Based on the interview data and survey responses, it seems that the majority of participants found the system easy to use and thought that most people would learn to use it quickly. However, there were some participants who found the system unnecessarily complex or very cumbersome to use. The majority of participants were pleased with the UI.

The SUS scores, as obtained from the final survey, can be seen in Table 4.4. To calculate the total SUS score, we used a formula where for the odd-numbered questions, we subtracted 1 from the scale position, and for the even-numbered questions, we subtracted the scale position from 5. The scores were summed and multiplied by 2.5 to obtain the total score of 68.33[35]. This indicates an average usability rating, as defined by Jeff Sauro [31].

Table 4.4: SUS Scores of Deep Dives

Question	P1	P2	P3	P4	P5	P6	Total
1	4	4	2	3	4	3	3.33
2	2	2	4	2	3	4	2.83
3	3	4	3	5	4	4	3.83
4	1	1	3	1	2	1	1.5
5	4	4	1	4	4	3	3.33
6	3	3	3	1	4	2	2.67
7	5	4	3	4	4	4	4
8	1	2	4	2	2	2	2.17
9	3	4	3	4	5	4	3.83
10	2	1	2	2	2	2	1.83
Totals	75	77.5	40	80	70	67.5	68.33

Chapter 5

Discussion

The results presented in this study reveal that there are several areas where the usability of AI-based writing and grading systems can be improved. As computer science education increasingly places value on technical writing skills, the use of automated systems has become more important to assist with grading and providing feedback that will help students expand their skills. The development of effective and user-friendly automated essay scoring systems can improve the learning experience for computer science students by providing them with more timely and consistent feedback on their writing assignments. Additionally, the incorporation of explainability and transparency techniques in these systems can help students understand how their writing is being evaluated and graded, which can lead to an improved understanding of artificial intelligence and increased AI literacy, all while improving their technical writing skills. These topics are important to be considered before adopting an AI-based grading tool, and this set of key considerations can be used to help develop more effective grading tools.

It is also important to note that different AI tools are more effective for different jobs, and there is a trade-off when deciding which type of tool to use. In educational contexts, several considerations come into play, ranging from the depth of feedback and complexity to the level of explainability of the algorithm.

On one end of the spectrum, when high accuracy is required, those systems prioritize solutions that are straightforward, clear, and explainable, as they have a larger impact on human

well-being. These solutions must offer a high degree of consistency and reliability, making them simpler to comprehend and manage.

On the opposite end of the spectrum, there are cases when accuracy is less important, such as when providing formative feedback. Those systems may prioritize solutions that are intricate, multifaceted, and exhibit variability, making them less straightforward to explain. These solutions, while potentially more powerful and versatile, may lack the simplicity and predictability associated with explainable systems.

Within the Deep Dives platform, the incorporation of instructor review as an obligatory step in the workflow allows for the adoption of slightly less explainable models. This strategic decision relies on the inclusion of human oversight within the process.

Overall, this spectrum represents a decision making process where choices are made based on the desired level of complexity, consistency, and explainability, with trade-offs considered depending on the specific context and objectives of the system.

5.0.1 RQ1: How do explainability and algorithm transparency techniques affect the overall usability and user experience of an AI-based essay feedback system?

The findings from reflexive thematic analysis and the System Usability Scale (SUS) scores provide insights into the impact of explainability and algorithm transparency techniques on the overall usability and user experience of the AI-based essay feedback platform.

Based on the theme of “Clarity of Feedback and Explanations”, participants, including both graders and students, highlighted the importance of clear and specific feedback that helps them understand how essays are being evaluated. In some instances, the AI system’s feedback

was considered vague or contradictory, leading to confusion among students. This suggests that improvements in the clarity and specificity of feedback and explanations could enhance the overall usability and user experience of the system.

The theme of "Impact and Actionability of Feedback" was influenced by participants' trust in the AI system. Gradual familiarity with the system and understanding its underlying algorithms led to increased trust over time. However, some participants expressed concerns about certain aspects of the feedback and grading criteria, which were perceived as subjective or not aligned with their own grading criteria. This indicates that further transparency and explainability of the AI system's evaluation process could enhance users' trust and increase their willingness to act upon the provided feedback.

In terms of usability, participants' feedback highlights the importance of clear and accessible information regarding the system's features, evaluation criteria, and expectations. The need for better organization of feedback, ease of finding explanations, and the inclusion of inline comments demonstrates a desire for more transparency in how the AI system operates and provides feedback. By addressing these usability concerns, the system can enhance transparency and explainability. Clear organization and accessibility of feedback help users understand how their essays are being evaluated and what specific areas they need to focus on for improvement. Inline comments can provide additional context and explanations, fostering the human-machine teaming approach by making the feedback more informative and actionable. More comprehensive information about the AI's expectations contribute to a clearer understanding of how the system operates, and helps users learn AI-decision making processes along the way.

Therefore, incorporating these enhancements in the system's usability not only addresses users' practical needs but also promotes transparency and explainability. Users can have a better grasp of the system's inner workings, the factors influencing their scores, and the

rationale behind the provided feedback, thus improving their AI literacy. This fosters a sense of trust and understanding, ultimately improving the overall user experience and the effectiveness of the system.

5.0.2 RQ2: How do graders perceive the integration of an automated essay feedback system into their grading process, and what are factors influencing their acceptance or resistance of automated feedback?

The survey results suggest that the incorporation of explainability and transparency techniques in an automated essay scoring system has an influence on the way graders used the system, as revealed by the findings from the reflexive thematic analysis and the System Usability Scale (SUS) scores.

Thematic analysis identified several themes related to the impact of explainability and transparency on human grading behavior. One theme that emerged was the influence of system explanations on graders' decision-making process. Clear and detailed explanations regarding the AI's assessment criteria and the factors contributing to the assigned scores helped align human graders' evaluations with the automated system. Gradual understanding of the algorithms and familiarity with the system's functioning empowered graders to make informed judgments, increasing their confidence in the automated scores. One SUS question assessed how confident participants felt using the system. The average score for this question was 3.83 out of 5, indicating a moderately high level of confidence and leaving room open for improvement.

Another theme that emerged from the thematic analysis was the influence of prior experiences

and expertise on human grading behavior. Graders with previous experience in automated essay scoring systems or a writing background demonstrated a deeper understanding of how the algorithm worked. Their expertise allowed them to critically evaluate the AI-generated scores, identify system strengths and limitations, and make informed decisions when overriding scores or providing additional feedback, thus highlighting the importance of improving users' AI literacy to increase their user experience with tools of this nature.

5.0.3 RQ3: What are the key components that constitute an effective automated essay scoring system, and how can they inform the development and assessment of reliable grading and feedback tools?

The SUS scores and thematic analysis provided insights into the system's usability and where its strengths and room for improvements lie. For example, participants highly appreciated the presence of small checks for specific aspects, such as word count objectives, which contributed to a user-friendly experience. The SUS question related to the system's ease of use received an average score of 4.6 out of 5, indicating a high level of satisfaction. This demonstrates that user-friendly features play a crucial role in enhancing the overall usability of the system.

The qualitative analysis revealed specific components that influenced user experience, such as clarity of feedback and explanations. Both graders and students expressed the need for more explicit and informative feedback. Participants desired clearer definitions, more examples, and a better understanding of how the system calculated scores for different rubric categories. This highlights the importance of providing comprehensive explanations and specific guidance to enhance the effectiveness of the feedback and grading process.

Furthermore, thematic analysis revealed the theme of trust in the automated scores. Participants gradually developed trust in the system through familiarity and understanding of its functioning. However, concerns were raised regarding certain components which were perceived as subjective or contradictory at times. This emphasizes the importance of algorithm transparency and continuous evaluation to enhance the system's reliability and foster trust among users.

In conclusion, an effective automated essay scoring system comprises key components such as clear and informative feedback, user-friendly features, transparent algorithms, and trust between users and the system. Enhancing the clarity of feedback, addressing user concerns, ensuring algorithm transparency, and refining user-friendly features are vital steps in developing reliable grading and feedback tools that enhance the overall usability and user experience of automated essay scoring systems. To help advance computer science education, attempts to build users' AI literacy by revealing certain decision making processes of the AI-algorithm is a highly rewarding and impactful practice. Continuous evaluation and user feedback are essential for iterative refinement and the development of effective tools in educational settings. During the evaluation process, specific questions should be asked about the ability to provide feedback to users, the transparency and explainability of the system, and the impact of the system on user learning and understanding.

5.1 Ethical Considerations

In order to protect the participants' rights and privacy, a number of ethical considerations were taken into account throughout the study. Before participating in the study, participants were sent an email that served as an informed consent form that explained the purpose of the study, the procedures involved, and potential risks and benefits of participating. The

recruitment email is available in A.7. In order to protect participants' privacy, their names and courses were not disclosed in any part of the study. Participants were told that their responses would be kept anonymous, although given the small and established group, it is possible that their identities could be revealed incidentally in the process. After each phase of the study, participants were given an opportunity to review and redact their comments before they were included in the findings of this study.

5.2 Limitations

This study is subject to a number of limitations that are necessary to address. First, due to the small selection of classes that utilized Deep Dives this semester, there were only a limited number of participants that were willing to participate, and thus their experiences may not be generalizable to a larger population. Furthermore, eight out of the nine participants were using Deep Dives in a computer science course, so this study does not fully address the use of this platform in a non-technical setting. This study only took place at one institution, which may limit the applicability of the findings to other settings. The results of this study were obtained over the course of a single semester. Thus, the long term impacts of using Deep Dives as a grading assistant is unknown. It is important to note that there have since been updates to Deep Dives that were not available when the study took place. The findings presented in this study may not be reflective of the latest version of the Deep Dives platform.

Additionally, throughout our study, we predominantly treated Deep Dives as a unified platform, overlooking the fact that it comprises a diverse array of underlying algorithms and functionalities. Our approach centered on evaluating the overarching impact and usability of Deep Dives as a grading assistant, considering the platform holistically. However, it is important to recognize that Deep Dives is a complex system, with various components working

together to perform specific tasks, such as automated essay grading, plagiarism detection, and feedback generation.

Despite these limitations, the results of this study can still serve as valuable starting points for researchers and educators interested in considering and developing AI-driven essay grading systems in their own contexts. The study's findings offer valuable initial perspectives and considerations for further exploration and can provide a foundation for future research in this domain.

Chapter 6

Conclusions and Future Works

In conclusion, this paper has presented a set of key considerations for evaluating the usability of AI-based essay grading tools. These considerations include, but are not limited to, clarity of feedback and explanations, impact and actionability of feedback and explanations, user understanding of system, trust in AI, issues and concerns, strengths of system, and the user interface.

Future research in this domain could delve into a more individualized analysis, dissecting each individual algorithm and piece of functionality that Deep Dives is composed of. By doing so, researchers could gain deeper insights into the strengths and weaknesses of each component, potentially uncovering opportunities for refinement and enhancement to optimize the overall performance of Deep Dives. Additionally, the key evaluation considerations proposed in this study hold significant potential for informing future research in the field of AI-based essay feedback systems and promoting design choices that improve users' technical writing while developing their knowledge of AI-driven systems. These considerations can serve as a valuable tool for researchers seeking to evaluate the usability and user experience of similar systems, as well as investigate the impact of explainability and transparency techniques on grading behavior and the development of reliable grading and feedback tools. Additionally, the list of considerations can be adapted and expanded to encompass other dimensions and variables specific to different contexts and target users. By applying this approach in future research, scholars can gain a deeper understanding of the effectiveness and limitations of AI-

based essay feedback systems, ultimately contributing to the improvement of these systems and enhancing their usability and educational value.

Bibliography

[1] Donald A. Norman. *The Design of Everyday Things*. Basic Books, Inc., 2002.

[2] Janet Davis. Design methods for ethical persuasive computing, 2009.

[3] BJ Fogg. Persuasive computers: perspectives and research directions, 1998.

[4] Carol Moser, Sarita Y. Schoenebeck, and Paul Resnick. Impulse buying: Design practices and consumer needs, 2019.

[5] Harry Brignull, Marc Miquel, Jeremy Rosenberg, and James Offer. Dark patterns-user interfaces designed to trick people, 2015.

[6] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. Dark patterns in proxemic interactions: a critical perspective, 2014.

[7] Renee Hobbs. Propaganda in an age of algorithmic personalization: Expanding literacy research and practice. *Reading Research Quarterly*, 55(3):521–533, 2020. https://doi.org/10.1002/rrq.301.

[8] H. Wang, C. Ma, and L. Zhou. A brief review of machine learning and its application. In *2009 International Conference on Information Engineering and Computer Science*, pages 1–4, 2009.

[9] Shweta Lamba, Preeti Saini, Vinay Kukreja, and Bhanu Sharma. Role of mathematics in machine learning. *SSRN Electronic Journal*, 2021.

[10] Jeremy Petch, Shuang Di, and Walter Nelson. Opening the black box: The promise

and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, 38(2):204–213, 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[12] Lauren E. Willis. Deception by design. *Harvard Journal of Law & Technology*, Volume 34:116–190, 2020.

[13] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[14] David Williamson, Xiaoming Xi, and F. Breyer. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31:2–13, 2012.

[15] Vivekanandan Kumar and David Boulanger. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5, 2020.

[16] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[17] Violet Turri. What is explainable ai? Carnegie Mellon University, Software Engineering Institute's Insights (blog), Jan 2022. Accessed: 2023-Aug-8.

[18] Margot E. Kaminski. The right to explanation, explained. *Berkeley Technology Law Journal*, 34:189, 2018.

[19] Duri Long and Brian Magerko. *What is AI Literacy? Competencies and Design Considerations*, page 1–16. Association for Computing Machinery, 2020.

[20] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution

and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[21] Emilee Rader, Kelley Cotter, and Janghee Cho. Explanations as mechanisms for supporting algorithmic transparency, 2018.

[22] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020.

[23] Aaron Springer and Steve Whittaker. Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Trans. Interact. Intell. Syst.*, 10(4):Article 29, 2020.

[24] Dikli Semire. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006.

[25] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. Generating diverse code explanations using the gpt-3 large language model, 2022.

[26] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions, 2021.

[27] Huanyu Bai, Zhilin Huang, Anran Hao, and Siu Cheung Hui. Gated character-aware convolutional neural network for effective automated essay scoring, 2022.

[28] Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. Formative essay feedback using predictive scoring models, 2017.

[29] Mohammad Alshammari, Rachid Anane, and Robert J. Hendley. Usability and effectiveness evaluation of adaptivity in e-learning systems, 2016.

[30] Luciana Freire, Pedro Arezes, and José Campos. A literature review about usability evaluation methods for e-learning platforms. *Work*, 41:1038–1044, 2012.

[31] Jeff Sauro. Measuring usability with the system usability scale (sus), 2011.

[32] Christina Kim Jacob Hilton Jacob Menick Jiayi Weng Juan Felipe Ceron Uribe Liam Fedus Luke Metz Michael Pokorny Rapha Gontijo Lopes Shengjia Zhao Arun Vijayvergiya Eric Sigler Adam Perelman Chelsea Voss Mike Heaton Joel Parish Dave Cummings Rajeev Nayak Valerie Balcom David Schnurr Tomer Kaftan Chris Hallacy Nicholas Turley Noah Deutsch Vik Goel Jonathan Ward Aris Konstantinidis Wojciech Zaremba Long Ouyang Leonard Bogdonoff Joshua Gross David Medina Sarah Yoo Teddy Lee Ryan Lowe Dan Mossing Joost Huizinga Roger Jiang Carroll Wainwright Diogo Almeida Steph Lin Marvin Zhang Kai Xiao Katarina Slama Steven Bills Alex Gray Jan Leike Jakub Pachocki Phil Tillet Shantanu Jain Greg Brockman Nick Ryder Alex Paino Qiming Yuan Clemens Winter Ben Wang Mo Bavarian Igor Babuschkin Szymon Sidor Ingmar Kanitscheider Mikhail Pavlov Matthias Plappert Nik Tezak Heewoo Jun William Zhuk Vitchyr Pong Lukasz Kaiser Jerry Tworek Andrew Carr Lilian Weng Sandhini Agarwal Karl Cobbe Vineet Kosaraju Alethea Power Stanislas Polu Jesse Han Raul Puri Shawn Jain Benjamin Chess Christian Gibson Oleg Boiko Emy Parparita Amin Tootoonchian Kyle Kosic Christopher Hesse John Schulman, Barret Zoph. Introducing chatgpt, November 30, 2022 2023.

[33] Teach with gpt, 2023.

[34] Packback ai ethics policy, 2019.

[35] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 1995.

[36] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101, 2006.

[37] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019. doi: 10.1080/2159676X.2019.1628806.

Appendices

Appendix A

User Study Documents

A.1 Pre-Study Survey Questions

1. For what other courses have you served as a TA?
2. How would you rate your level of confidence in your ability to consistently grade with the same level of harshness?
3. How would you rate your level of confidence in grading consistently with other graders of the same course?
4. Have you graded writing assignments before?
5. If so, was this from a past course or this one?
6. Estimate how long it takes you to grade a one page paper
7. What part of grading writing assignments takes the longest and why?
8. What parts of grading rubrics do you typically put the most effort into?
9. What other AI systems have you interacted with before?
10. Please rate your level of knowledge of artificial intelligence and machine learning algorithms on a scale from 1 to 5

11. Have you ever used an automated grading tool before?
12. If so, please state the name of the tool, if known.
13. What positive experiences do you remember having when using the tool?
14. Were there any issues or confusions you experienced when using this tool? Please explain
15. What was the automated system doing that was obvious and what was the system doing that was difficult to interpret?
16. Have you used Packback Deep Dives before?
17. What do you expect to be the advantages of Deep Dives?
18. What worries do you have about using Deep Dives to assist with grading?

A.2 Midpoint Survey Questions

1. Roughly how many individual Deep Dives responses have you graded so far?
2. How many different Deep Dives assignments have you been a part of grading?
3. Please describe your initial impressions of using Deep Dives
4. Have you encountered any problems or issues with the platform?
5. If so, please describe the issues you have faced
6. Is there anything that you are still confused about when using the tool?
7. If so, please describe
8. What have you found to be the advantages of using the tool so far?
9. What have you found to be the disadvantages?
10. So far, how does your grading efficiency when using Deep Dives compare to grading without the assistance of an automated tool?
11. So far, how does the time it takes to grade an assignment using Deep Dives compare to grading without the assistance of an automated tool?
12. How much do you trust the accuracy of the automated grades Deep Dives provides?
13. Do you typically go back and double check to see if the system generated grades seem fair?
14. How often have you had to override the scores provided by Deep Dives?
15. How strong of an understanding do you think you have of how the automated grading system works?

16. Please briefly describe how you understand the grades are generated

A.3 Post-Survey Questions

System Usability Scale (from 1 to 5)

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system

11. How would you rate your level of confidence in grading consistently and with the same level of fairness throughout this course?
12. How would you rate your level of confidence in grading consistently with the other graders of this course during this semester?
13. On average, how long was an individual Deep Dives response?

14. On average, how long did it take you to grade a single Deep Dives assignment?

15. What part of grading Deep Dives assignments took you the longest and why?

16. What part of the Deep Dives rubric did you typically put the most effort into when grading?

17. How much feedback did you typically give for a single Deep Dives response?

A.4 TA/Instructor Interview Questions

1. Clarify background with AI, experience with automated grading, etc. (specific to each participant)
2. Grading experience
3. Knowledge of AI
4. Previous interactions with AI systems
5. Clarify anything from Pre and Midpoint survey that needs to be clarified (specific to each participant)
6. What was your overall experience with using the platform?
7. What did you find to be the advantages of using the tool?
8. Disadvantages?
9. Was there anything that was difficult to understand about the tool?
10. Was there anything that you wanted to be able to do that the system couldn't provide or didn't support?
11. How did the use of Deep Dives change the way you grade writing assignments?
12. Is there anything you would change about the system?
13. Would you prefer to use this product over the way you traditionally graded
14. Why?
15. How much time, if any, do you feel you saved with grading?

16. Did using this tool save you effort? Why or why not?
17. Do you feel using this tool made grading more efficient? Why?
18. Did you have an easy time finding explanations for automated grades?
19. How easy was it to override grades?
20. (For professor) how easy was it to adjust the grading rubric and weight of each category?
21. How well did you understand how certain grades were made?
22. Can you describe to me how you think the algorithm works?
23. How did you come to that conclusion?
24. What was the system doing that was obvious/easy to interpret?
25. What was the system doing that was confusing/difficult to interpret?
26. Did you ever have questions about the automated grades? Were you able to easily find the answers?
27. Explain how you think each of these scores were generated
    - Wordcount and depth
    - Grammar and mechanics
    - Flow and structure
    - Research quality
    - Formatting
28. Did students ever question the scores they were given?

29. If so, which ones?

30. Did you trust the automated scores? Or did you feel like you had to go through and check its work?

31. How often did you choose to override scores?

32. Why did you choose to override those scores?

33. Did the automated grader typically grade higher, lower, or the same as you would have?

A.5 Student Interview Questions

1. What was your overall experience with using the platform?
2. What do you think are the advantages of Deep Dives?
3. Disadvantages?
4. Would you prefer to write using Deep Dives over other alternatives (write in Word Doc, submit through Canvas)? Why?
5. Did you have an easy time finding Deep Dives provided feedback?
6. Explain how you think each of these scores were generated
   (a) Wordcount and depth
   (b) Grammar and mechanics
   (c) Grammar and mechanics
   (d) Flow and structure
   (e) Research Quality
   (f) Formatting
7. How did you come to that conclusion?
8. What feedback did you receive that was clear to you? Why was it clear? What did you do in response? Did you change your writing in any way?
9. What feedback did you receive that was difficult to interpret? Why was it confusing? What, if anything, did you do in response to that feedback?

10. Can you give me specific examples of feedback that made you go back and change what you wrote?

11. Can you give me specific examples of feedback that didn't make you go back and change your writing?

12. Was there a difference in the automated feedback you received while writing vs. after it was graded? If so, what was different about it?

13. Was it always clear to you what you had to fix in order to improve your automated scores? If there were times where it wasn't, what would you do in response?

14. After receiving grades back, did you ever have questions about the automated scores? If so, were you able to easily find the answers? How?

15. Did you typically go back and look at the score breakdown after your papers were graded?

16. If so, was it clear to you which grades came from the automated system and which grades came from the professor/TAs? How did you know/guess where the grade was coming from?

A.6 Focus Group Questions

1. What were your expectations of Deep Dives prior to using the product?
2. What were your overall impressions of Deep Dives?
3. What did you like about Deep Dives and why?
4. What didn't you like about Deep Dives and why?
5. How easy was it to give and understand feedback? Was the AI generated feedback sufficient for you? - Why?
6. Was it clear to you what AI feedback the students/instructors were seeing? -Why?
7. Did you encounter any problems with the UI? - What?
8. Were there any major learning curves you had to get over when first using the product? Explain?
9. Were you able to easily find explanations for the automated grades?
10. Were some rubric categories more clear in how they calculated the grades than others? If so, which ones? - Why?
11. Do you feel you had a good understanding of how the automated system was grading/how you were being graded? Why or why not?
12. Starting with TAs/Instructors, could you all briefly describe how you understood how the automated system was grading?
13. Students, do you agree or disagree with these descriptions? Why or why not?
14. Did you trust the automated grades? Why or why not?

15. Would you all prefer to use this product over the way you traditionally graded/wrote papers? - Why?

A.7 Recruitment Email

Hi,

My name is Erin Hall, and I am currently pursuing my master's degree in computer science, with a focus in HCI, usability engineering, and AI. I am working with Dr. Mohammed Seyam and Dr. Dan Dunlap to conduct a usability study of Packback Deep Dives this semester. As part of our effort to evaluate Packback Deep Dives, I am hoping you will be able to participate in a few surveys and an interview to get your perspectives and experiences with using the platform. As a Instructor/GTA of a course using Packback Deep Dives this semester, your views are extremely important and will contribute to academic research related to the evaluation of similar AI-based educational platforms. I hope you might have about an hour to answer some questions about your experiences with the platform, and about 10-15 minutes to complete 3 related surveys. Please let me know as soon as possible if you would be willing to participate.

Here are a few points to be aware of:

1. You will be sent out three digital surveys throughout the course of the Fall 2022 semester. The surveys will be conducted through Google Forms and sent over email. The surveys should take about 10-15 minutes each to complete.

2. The interview portion of the study will last for about an hour. The interview will be held over Zoom and will take place towards the end of the semester, after all Packback-related assignments have been completed.

3. With this in mind, it would be valuable to document your impressions of using the platform along the way to help remember your experiences from earlier in the semester.

4. I (Erin Hall erinehall@vt.edu) will be conducting the interview. I am not connected to the Packback in any way outside of this research.

5. The interview will be more like a conversation than asking questions on a structured guide. Prior to the interview, you will receive an informed consent form with details about your decision to consent or not, and I will be happy to answer any and all questions about the questions and collection of this information.

6. Participation in this interview is your choice, and I will not share your decision with others. You will have an opportunity to redact your comments and to consider your answers before any of your comments are shared with other researchers and/or employees of Packback.

7. I will do my best to keep all comments anonymous, but because of the nature of the small and established group, it is possible that your identity could be obvious to some and/or revealed incidentally in the process. You should keep that in mind when you agree to participate and as you answer questions. This will be explained if and when I discuss the informed consent prior to the interview.

With best regards,

Erin Hall

Appendix B

Reflexive Thematic Analysis Results

B.1 Clarity of Feedback and Explanations

Table B.1: Clarity of Feedback and Explanations

Graders	**Students**
• Sometimes students did/didn't get full credit and it wasn't clear why • Didn't always know what it was looking for • Wanted a better understanding, especially of flow and structure • Didn't always understand the "see why" blurb • Didn't always understand how those subcategory numbers were calculated • Explanations didn't always give enough information • Hedge words wouldn't say which specific words and where they were located in the essay	• Would complain early on for readability when only a few sentences were written • Some feedback was vague • AI complaining about minor things wasn't a big deal because TAs helped reduce the point loss on ridiculous things • Likes the word count objective • Clear feedback: Easy solutions like adding something or taking something away • Confusing feedback: Hedge words, Flesch Kincaid score • Sometimes feedback was contradictory • Would get dinged for Flesch Kincaid score, which didn't make sense because it was a college course • Likes the small checks Deep Dives has • Liked the dynamic feedback while writing • Deep Dives would give feedback but wouldn't tell you how to fix it • Didn't know where the issues were happening

Table B.2: Clarity of Feedback and Explanations (Continued)

Graders	**Students**
• Didn't have examples of what flow and structure category was looking for • Didn't know what the AI was looking for but knows what they would've looked for • Wishes there were more clear definitions of what they were looking for • Deep Dives was vague, and more information would be better • Sometimes point deductions were obvious because of students writing • Pretty clear where grades were coming from • Wishes there was more feedback and that they could see specific examples of what was wrong like the students could • Couldn't always answer student questions when they reached out	• Broad and vague feedback about your writing as a whole was confusing and hard to fix because it wasn't specific enough • Formatting and saying their paragraphs were too long were confusing feedback because it complained even when paragraphs were a sentence long • Realized the comments were always the same–it would just hide the ones that didn't apply • Usually it was clear what to fix to improve scores • Sometimes feedback was contradictory • Deep Dives would give feedback but wouldn't tell you how to fix it • Word count, number of sources, grammar issues, and credibility were all clear pieces of feedback • Clearest feedback was most basic–either you have it or you don't • Broad and vague feedback about your writing as a whole was confusing and hard to fix because it wasn't specific enough • Realized certain cheats after writing a few essays • Could see a progress circle but not the number of points they had

B.2 Impact and Actionability of Feedback and Explanations

Table B.3: Impact and Actionability of Feedback and Explanations

Graders	**Students**
• Would adjust unclear points • Took away a lot of tedious aspects–word count, plagiarism, etc. • Would just glance at everything to make sure it was accurate • Double-checked first few times and found errors, so felt always needed to double-check • Completely stopped looking at reference grades because they were always wrong	• Would ignore AI feedback when they couldn't figure out how to fix it • Sometimes feedback was difficult to act on • Feedback that required more effort they would typically leave alone • Certain things the AI complained about weren't reasonable enough for them to pay attention to • Usually it was clear what to fix to improve scores • Would sometimes change writing in response to feedback • Low effort changes where feedback was specific would make them change their writing • Feedback that required more effort they would typically leave • Realized certain cheats after writing a few essays • Usually left their writing as is because they couldn't figure out how to get the confusing feedback errors to go away • When they thought something was weird, would just ignore the AI • When they thought the standard was too high, would ignore the AI

B.3 Understanding of System

Table B.4: Understanding of System

Graders	Students
• Previous experience with similar tools helped them figure out how the system worked • Grammar and mechanics use NLP, recognizes grammatical mistakes • Structure has specific rules like words per paragraph, number of paragraphs, etc. • AI grader usually graded the same as they would've • Deeper understanding would help determine which categories they trusted more • Would be easier if students couldn't see concrete numbers • AI and writing background helped them understand how the algorithm worked • Understood scores but didn't always agree with them • There wasn't information available to figure out how scores were calculated • Better understanding of the algorithm would help ensure they were doing a good job as a grader	• Helped them figure out how close they were to being done • Score would naturally improve as they wrote more • Flow and structure was most confusing to understand • Knew how algorithm worked from guesses, trial and error, and discussing with professors and other students • Thinks depth uses AI that goes through published articles and bases its parameters on what it reads • Grammar algorithm is a parser with a checklist of grammar rules • No idea how flow and structure was calculated, played around with it a lot and best guess is it has to do with the amount of words in a sentence

Table B.5: Understanding of System (Continued)

Graders	**Students**
• Not sure how formatting and presentation was calculated • Research and citations would deduct points if they were missing author, date, etc. • Assumes algorithm is based on reading papers and noticing patterns • AI and writing background helped understand how the algorithm worked • Understood how it worked based on a little information Packback gave and reading and picking up patterns • Having a better understanding of how auto grades were generated would be constructive • Didn't know what the students could see • Assumed students saw the same feedback they did • Low scores indicated an error	• Sources with more foot traffic were counted as more credible • Knew how the algorithm worked because some pieces of feedback were clear, and also a lot from assumptions, playing around with it, and talking to people • Thinks if it ends in .gov, it might be more credible than .net and .com • AI was helpful since Deep Dives flags websites as unreliable if they are too partisan • Usually, sources with DOI numbers were highly credible • Thinks there might have been a rubric for each Deep Dives evaluation on the canvas page • Thinks for the highly credible ones, they have a static list of specific domain names • Wasn't always obvious when TAs overrode grades • Just assumed when professors overrode grades

B.4 Trust in AI

Table B.6: Trust in AI

Graders	Students
• Got more comfortable with the algorithm as time went on • Didn't put too much faith in the AI • Double-checked scores to make sure they were fair • Trusted grammar and mechanics the most • Trusted flow and structure the least • AI grader usually graded the same as they would've • Trusted it more and less over time because they could see how it worked, but didn't always agree with how it graded • Always had to go back and check • Got more comfortable with the algorithm as time went on • Trusted it more as time went on because they learned what they could trust • Knew what to expect over time	• Usually didn't question the AI • Usually got better grades than Deep Dives made it seem like • Figured out Deep Dives' quirks over time

Table B.7: Trust in AI (Continued)

Graders	**Students**
• Biggest learning curve was getting used to the AI • Had to tailor expectations to what AI would actually offer • Never bumped down on AI grades • Didn't think it was overly harsh • Would flag grammar things that weren't actually issues • Trust depends on category and grade–didn't trust low grades • Overrode grades for almost every student • Didn't trust it less over time but assigned it a different role–to assist grading and not replace it • Didn't override scores often • Trust depends on category and grade–didn't trust low grades • Sometimes it graded too harsh, sometimes not harsh enough • Felt like it was harsh • Didn't think it was overly harsh • Trusted it more once they noticed it graded like they would • Would override scores when they didn't trust grades and didn't know what was going on • Didn't override scores often	• Learned not to take Deep Dives scores at face value • Thinks the AI was generally fair

B.5 Issues and Concerns

Table B.8: Issues and Concerns

Graders	Students
• Questions if it is fair to put students in a box and force them to write in a certain way • Flow and structure is pretty subjective • Would bump up flow and structure scores when students got deducted points for flowery writing • Students questioned flow the most • Forced students to write unnaturally • Sometimes had a differing opinion with AI, so didn't know if they should change it • Would override when they didn't agree with score • Hesitant to override because didn't want their own bias to bleed in • More boring papers were more likely to get full credit	• People had different experiences because Deep Dives are telling you one thing, but TAs are grading for another • Realized there were certain cheats • Sometimes feedback was contradictory • Wants to see all pieces of feedback at once • They can't have a discussion with AI like they could with a professor; wished for more information • The software prohibited their own writing process and wanting to expand on certain thoughts • AI marks some sources as somewhat credible that they thinks should be highly credible • Realized that the way it rates credible sources is dubious at best, so wouldn't pay as much attention to it • Issues come when there are grammar rules that have exceptions • It made you dumb it down • There were a lot of errors • Didn't know where issues were • Said paper looks choppy when it was only two sentences long • 300-word paragraph with 10 sentences was too short, but 4-sentence paragraph was too long

B.6 User Interface

Table B.9: User Interface

Graders	Students
• UI was easy to use • Finding feedback was easy • Feedback was somewhat hidden • 3 or 4 clicks to see what was being flagged • UI was sometimes weird • Sometimes couldn't find the explanations • Had to manually transfer grades to gradebook • Feedback isn't all on one page • Easy to provide feedback	• Everything was easy to find within the UI • Usually left their writing as is because they couldn't figure out how to get the confusing feedback errors to go away • Font was sometimes a bit small • Couldn't figure out how to see more than two pieces of feedback

B.7 Areas of Improvement

Table B.10: Areas of Improvement

Graders	Students
• Wants inline comments • Would flag grammar things that weren't actually issues • Wants a better plagiarism check • Wishes they could see what the AI was asking the students to do	• Wishes they could see all pieces of feedback at once so they could fix the easier ones first • Wishes Deep Dives gave more information because they can't have a discussion with AI like they could with a professor