Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #46211

Leveraging NLP for Classifying Student Ethical Responses in an Engineering Narrative Game

Ms. Tori N. Wagner, University of Connecticut

Tori Wagner is a doctoral student at the University of Connecticut studying Engineering Education. She has a background in secondary science education, playful learning, digital game design, and Natural Language Processing.

Dr. Daniel D. Burkey, University of Connecticut

Daniel Burkey is the Associate Dean of Undergraduate Programs and Associate Professor in the Department of Chemical and Biomolecular Engineering at the University of Connecticut. He received his B.S. in chemical engineering from Lehigh University in 1998, his M.S. and Ph.D. in Chemical Engineering from the Massachusetts Institute of Technology in 2000 and 2003, respectively, and his M.A.Ed. In Educational Psychology with a specialization in Research Methods, Measurement, and Evaluation in 2023.

Leveraging NLP for Classifying Student Ethical Responses in an Engineering Narrative Game

This work-in-progress explores the application of pre-trained, open-source transformer models designed to run efficiently on local hardware for natural language processing (NLP) in classifying student short-answer responses within the context of the narrative-based engineering ethics game/assessment, *Mars! An Ethical Expedition* (*Mars!*). Building on the contemporary learning theory of situated cognition and concepts of seamless (stealth) assessment, the game immerses students in decision-making scenarios tied to ethical dilemmas on a Mars settlement, encouraging context-dependent ethical reasoning. The primary focus is on analyzing the justifications students provide for their in-game decisions using NLP-based text analytics. Traditional ethical reasoning assessment tools, such as the Engineering Ethical Reasoning Instrument (EERI), have been critiqued for their limitations in capturing in-situ ethical decision-making. In response to these limitations, *Mars!* was developed to provide a rich, narrative-driven environment that allows for a more context-sensitive assessment of students' ethical reasoning as they engage with complex, first-person dilemmas.

We propose using transformer-based machine learning techniques to analyze student responses, with a primary focus on assessing response completeness. The completeness classifier categorizes responses as irrelevant or incomplete, partial, or complete, providing instructors with a scalable method to evaluate student engagement with ethical dilemmas. Beyond completeness, models will categorize student justifications based on perspective (e.g., first-person vs. third-person reasoning), motive (e.g., self-interest vs. social good), and reasoning approach (e.g., strict rule application vs. situated reasoning). These additional classifications are designed to support future research into how students frame and justify ethical decisions. By evaluating these responses at scale, the study aims to develop more efficient and accurate instructor-friendly tools for assessing ethical reasoning in authentic, first-person contexts.

Initial results suggest that locally deployed transformer models for text classification may supplement quantitative ethical reasoning assessments like the EERI by providing additional nuanced analysis of student ethical judgments. This project contributes to a growing body of research on the use of text analytics for formative assessment in engineering ethics education, with implications for enhancing student learning and promoting ethical decision-making in professional engineering contexts.

Introduction

Ethical reasoning is a critical competency for engineers, as their decisions often carry profound societal, environmental, and safety implications. Traditional assessments of ethical reasoning, such as the Defining Issues Test (DIT) [1] and the Engineering Ethical Reasoning Instrument (EERI) [2], are modeled on Kohlberg’s justice-based moral development framework [3]. While these assessments provide quantitative measures of ethical judgment, they often fail to capture the complexity and context-dependence of ethical decision-making in real-world engineering practice.

A key limitation of these static, principle-based assessments is that they emphasize abstract reasoning over situated, in-the-moment ethical decision-making. Engineers do not make ethical choices in isolation; rather, they navigate high-pressure, context-sensitive environments where multiple competing values—such as safety, efficiency, and financial constraints—must be considered [4]. Traditional ethical reasoning instruments are often detached from these realities, assessing responses in third-person, hypothetical formats rather than first-person, immersive scenarios [5]. To address these limitations, there is a growing recognition of the need for assessment tools that evaluate ethical reasoning in first-person, authentic contexts.

Mars! An Ethical Expedition (*Mars!*) is a narrative-driven, game-based intervention designed to assess and cultivate situated ethical reasoning in engineering students [6]. Unlike traditional instruments that rely on hypothetical, multiple-choice responses, *Mars!* places students in first-person, high-stakes ethical dilemmas within a simulated Mars settlement. Players must navigate challenges such as resource allocation, environmental safety, AI failures, and professional responsibility, making choices that impact their team, mission, and long-term sustainability. These dilemmas are intentionally designed to align with ABET accreditation Criterion 3, Student Outcome 4, which emphasizes the ability to recognize ethical and professional responsibilities while considering the global, economic, environmental, and societal impacts of engineering solutions [7]. Traditional ethics assessments can struggle to capture this complexity, particularly in large lecture courses where in-depth role-playing or case studies are difficult to implement at scale. *Mars!* provides an alternative by embedding ethical decision-making into a narrative-driven experience, allowing students to engage with multifaceted dilemmas while enabling scalable assessment through natural language processing (NLP). The game is easy to use and was designed with accessibility in mind, offering features such as multiple input options (mouse, keyboard, game controller), dyslexic-friendly font settings, subtitles, and screen reader support for non-voice-acted text. Currently, student responses must be downloaded, compiled, and analyzed by a researcher as the transformer models continue to be refined. However, the long-term goal is to integrate NLP seamlessly within the game, enabling real-time analysis of student responses. Responses, completeness scores for each question, and an overall score will be automatically compiled and reported to instructors, making large-scale assessment feasible while preserving the depth of open-ended ethical reasoning. This study does not aim to replace instructor evaluation but rather to complement it by providing NLP-based classification of individual student responses. While the model assesses ethical reasoning at the individual level, its most reliable and useful application is in aggregating data at the classroom level. This allows instructors to identify broad engagement trends and support formative assessment and discussions without relying solely on manual evaluation. *Mars!* can serve as a valuable starting point for in-class discussions, where small or large group conversations can further unpack ethical dilemmas and reasoning strategies. In this way, *Mars!* functions both as a standalone assessment tool and as a springboard for deeper discussions, complementing other assessments in engineering ethics education while providing additional evidence for ABET accreditation.

Natural Language Processing (NLP)

This study explores how pre-trained, open-source transformer models can be used to classify student responses in *Mars!* using text analytics techniques. Specifically, it examines how student justifications for their ethical decisions can be categorized based on perspective (first-person vs. third-person), motive (self-interest vs. social good), reasoning approach (strict rule application vs. situated reasoning), and completion (depth of elaboration). To accomplish this, the study employs a combination of dictionary-based and transformer-based NLP models to automate the qualitative analysis of ethical reasoning, allowing for a more scalable and systematic approach to evaluating student responses.

To better understand where *Mars!* fits within the landscape of text analytics, it is helpful to briefly examine the evolution of NLP. The development of NLP has progressed through several key transformations, shifting from rule-based and statistical methods to deep learning-driven models that utilize vast datasets for language understanding and generation. This evolution can be categorized into four major stages: Statistical Language Models (SLMs), Neural Language Models (NLMs), Pre-trained Language Models (PLMs), and Large Language Models (LLMs)[8]. Early SLMs relied on probability-based predictions but struggled to capture meaning beyond short phrases. The shift to NLMs introduced word embeddings (Word2Vec [8]) and Recurrent Neural Networks (RNNs), which allowed computers to process text with some awareness of word order and context. Later models, like Long Short-Term Memory networks and Gated Recurrent Units, improved this capability but still had trouble understanding longer passages of text [9]. A major breakthrough came with transformers, which essentially allowed models to analyze entire sentences at once rather than one word at a time. This led to the development of PLMs, such as BERT, RoBERTa, DeBERTa, and MPNet, which improved text classification and interpretation [9]. Other models, such as T5 and BART, were designed to handle more complex tasks like summarization and translation [9]. While both PLMs and LLMs are trained on vast datasets and leverage the transformer architecture for natural language understanding, they differ primarily in scale, computational requirements, and application focus. LLMs, such as GPT-4 and Open LLaMA, process vast amounts of text and generate human-like responses with little down-stream training [9]. However, their size and complexity require significant computing power and often rely on cloud-based services, making them less practical for packaging directly within a game for live, real-time analysis.

Given the computational limitations and privacy concerns associated with massive LLMs, this study employs smaller, locally runnable, more secure PLMs rather than full-scale LLMs with the goal of eventually deploying the trained PLMs in-game for live assessment of qualitative game data.

Situated Cognition and Situated Learning

The Situated Cognition framework [11] posits that learning and reasoning are fundamentally shaped by the context in which they occur. Rather than being a purely abstract cognitive process, ethical reasoning is embedded in specific social and environmental conditions, requiring individuals to interpret problems within real-world constraints [12]. Within engineering

education, ethical reasoning is often assessed through hypothetical case studies that present generalized moral dilemmas in an abstract format. While these methods allow for structured evaluation, they often fail to capture the complexity of in-the-moment ethical decision-making that engineers face in professional practice. Ethical challenges rarely present themselves as clearly defined problems with pre-determined solutions; instead, they unfold in dynamic, high-stakes environments where individuals must interpret incomplete information, respond to unforeseen consequences, and balance competing priorities [13]. To address these limitations, *Mars!* is designed as a first-person, narrative-based game that immerses students in real-time ethical dilemmas. By placing students in authentic, problem-solving situations, *Mars!* encourages ethical reasoning that reflects the complexity and ambiguity of professional decision-making, rather than requiring students to apply pre-existing ethical frameworks in a detached, theoretical manner.

Playful Learning and Stealth Assessment

The integration of game-based learning into ethics education builds on research suggesting that playful environments encourage deeper engagement and more authentic decision-making [14]. Narrative-driven games like *Mars!* provide students with interactive, immersive experiences that require them to make ethical choices within realistic, high-pressure scenarios, rather than simply reflecting on ethical principles in hindsight [15]. One of the key advantages of game-based learning is its ability to place students in decision-making roles, encouraging them to grapple with ethical dilemmas in real time, rather than passively analyzing hypothetical case studies.

While the term "game" often implies winning, losing, and competitive strategy, *Mars!* functions more like an extended case study role-play than what one might consider a traditional game. It is not designed around competition or "right" answers, but rather to capture and analyze students' ethical reasoning as they navigate ambiguous scenarios. Choices in *Mars!* have consequences within the narrative, but students are not given completely open-ended agency; instead, they must choose between ethically complex options, all of which are defensible yet imperfect. This design encourages students to reflect on ethical trade-offs rather than simply selecting the "correct" response. Games exist along a spectrum of agency and engagement, shaped by design goals and player interaction. While students have some influence over the narrative through their ethical choices, their agency is constrained within structured dilemmas that present predefined yet morally complex options. Engagement comes not from open-ended decision-making, but from navigating ethically ambiguous scenarios and reflecting on the consequences of their choices within the game's framework.

A major advantage of this approach is the ability to integrate stealth assessment—a method that captures student learning and decision-making without disrupting engagement [16]. Traditional assessments, such as surveys and standardized ethics tests, often prompt students to rationalize decisions after the fact, rather than examining how ethical reasoning unfolds in real-time. By contrast, stealth assessment records decision-making processes within gameplay itself, providing a more naturalistic measure of ethical reasoning [17]. Because students are actively engaged in ethical problem-solving, their responses reflect situated, context-dependent reasoning rather than detached, theoretical reflection.

Methods

This research is designed as a qualitative text analytics study, leveraging machine learning models to classify student responses to ethical dilemmas in *Mars!*.

Participants

The study involves 398 engineering students enrolled in a first-year foundations of engineering course, where *Mars!* is used as a game-based ethics intervention. Over 12 weeks, students play through one chapter per week, engaging with narrative-driven ethical dilemmas that require real-time decision-making. Participants provide open-ended justifications for their choices, which serve as the primary data source for NLP-based analysis. Student responses are anonymized before NLP classification, and no personally identifiable information is processed.

Ethical Dilemma

The dilemma chosen to train the initial text classification models from the *Mars!* story was *Chapter 5: Infection* where players must decide whether to prioritize immediate rescue or adhere to strict quarantine protocols. In this scenario, Jonathan, a trusted assistant, is trapped in the airlock with an injured, angry dog he found wandering the Martian surface. The settlement has a mandatory 24-hour quarantine policy for anything that has been exposed to the outside environment, especially unknown life forms. However, Jonathan has already been bitten by the dog, and his condition may worsen if left untreated. The protagonist, acting as the Commander, must choose between two difficult options:

1. Break quarantine to immediately bring Jonathan inside for medical attention, ensuring he gets treatment as soon as possible.
2. Enforce the 24-hour quarantine, keeping Jonathan and the dog in isolation until it is confirmed safe to bring them inside.

The specific free-response question students are asked to answer after making their choice is:

"Describe the worst and best possible outcomes of your decision. Tell us about your thinking in regard to your decision and how considering the worst case affects your thinking."

The responses are analyzed using NLP models, which classify student reasoning based on completeness (e.g., partial or fully developed responses), perspective (e.g., first-person vs. third-person framing), motive (e.g., self-interest vs. social good), and reasoning approach (e.g., strict rule-following vs. context-sensitive decision-making).

Text Classification

Text classification, a fundamental task in NLP, was used to categorize student responses into predefined labels. In supervised text classification, a labeled dataset was used to train a machine learning model that learned to identify patterns distinguishing different categories. Once trained, the model's performance was evaluated to ensure reliability and validity before being applied to new, unlabeled data for large-scale assessment.

For this study, text classification models were trained to categorize student responses according to a predefined rubric assessing completeness, perspective, motive, and reasoning approach (see Table 1).

Table 1: Initial Rubric for Open-Ended Response Classification

Category	Completeness	Perspective	Motives	Reasoning Approach
0	Irrelevant or Incomplete: The response did not sufficiently address the ethical dilemma or was off-topic.	Neither: The response did not clearly indicate a specific perspective.	Neither: The response did not clearly indicate selfish motives or social good.	Neither: The response did not clearly indicate a strict rule-based or situated reasoning approach.
1	Partial Answer: The response engaged with the ethical dilemma but lacked depth, reasoning, or clear justification.	Third Person: The response was framed from an external perspective, focusing on the experiences, feelings, and actions of others.	Selfish: The response was primarily concerned with personal gain, benefit, or advantage, with little regard for others.	Strict Rule Application: The response focused on following predefined rules, laws, or principles without considering context.
2	Complete Answer: The response fully engaged with the ethical dilemma, providing a well-reasoned and justified explanation.	First Person: The response was framed from the player's own perspective, focusing on personal experiences, feelings, and actions.	Social Good: The response was primarily concerned with the well-being of others or the greater good of the community.	Situated Reasoning: The response considered the specific context, nuances, and situational factors in the ethical dilemma.

Four separate text classification models were trained to classify responses based on completeness, perspective, motive, and reasoning approach. Prior to training, the data underwent preprocessing steps to standardize text format and remove noise. The dataset was then split into training, test, and validation sets, with the training set used for model training, the test set reserved for hyperparameter tuning and model selection, and the unseen validation set for assessing generalizability (see Figure 1).

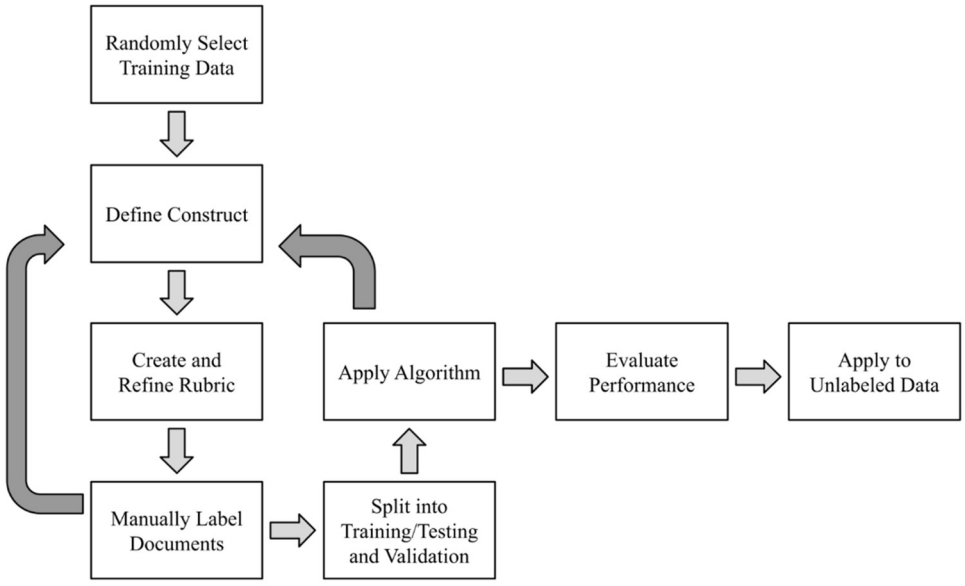

Figure 1: Supervised Text Classification Overview

For text classification, Microsoft's DeBERTaV3-small [18], a transformer-based model, was selected due to its strong performance in capturing contextual meaning and semantic relationships, as well as its small size and efficiency, making it well-suited for deployment in resource-constrained environments. The model was trained separately for each rubric category, taking preprocessed student responses as input and generating a probability distribution across the three predefined rubric levels (0, 1, 2). Model performance was evaluated using accuracy, precision, recall, and F1-score (the harmonic mean of precision and recall) on the test set to ensure reliability.

Preliminary Findings

The initial implementation of NLP models for ethical reasoning classification demonstrates promising results. The models successfully categorize student responses based on completeness, perspective, motive, and reasoning approach, enabling a more scalable and nuanced analysis of ethical reasoning patterns.

Completeness Classifier

The completeness classifier is of particular interest to classroom instructors, as it serves as the primary indicator of whether students are meaningfully engaging with the ethical dilemmas presented in *Mars!* and demonstrating ABET Student Outcome 4: the ability to recognize ethical and professional responsibilities and make informed judgments [7]. The Completeness Classifier evaluates the depth and quality of student justifications. Responses are classified into three levels:

Irrelevant or Incomplete (0) – indicating a lack of substantive engagement with the dilemma.

Partial Answer (1) – where students address the ethical issue but provide limited reasoning or justification.

Complete Answer (2) – demonstrating a well-reasoned, fully developed ethical justification.

Because the primary goal of *Mars!* as an assessment tool is to provide instructors with a scalable method of evaluating ethical reasoning, completeness scores offer the clearest measure of whether students are successfully articulating their thought processes.

The completeness classifier demonstrated strong initial performance, achieving an accuracy of 0.9167 and an F1 score of 0.9162, indicating high reliability in distinguishing between irrelevant/incomplete (0), partial (1), and complete (2) responses (See Table 2 and Figure 2). These results suggest that the model effectively identifies the depth of student engagement with ethical dilemmas, successfully capturing whether responses fully articulate reasoning, provide some justification, or fail to meaningfully address the question.

Table 2: Transformer-Based Completeness Classifier Metrics:

Accuracy	Precision	Recall	F1 Score
0.9167	0.9165	0.9167	0.9162

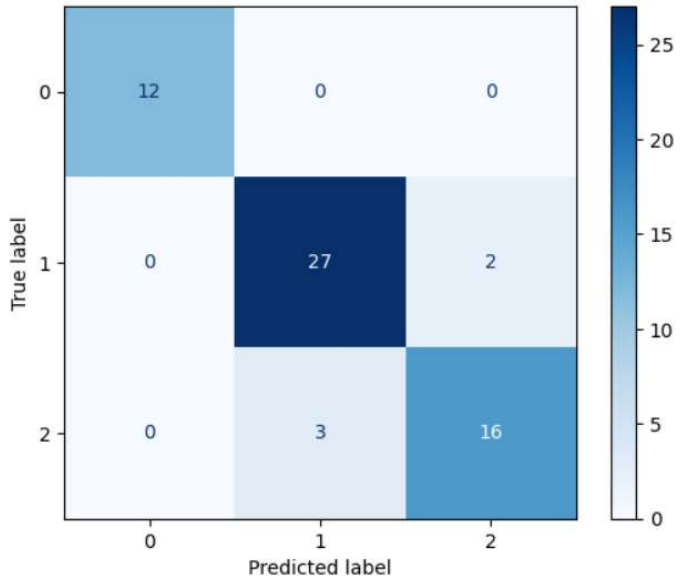

Figure 2: Confusion Matrix for Completeness Classifier Predictions

However, while overall performance is promising, early analysis of misclassified responses indicates that the model occasionally struggles to distinguish between partial (1) and complete (2) responses, particularly when students provide concise but well-reasoned justifications or lengthy but superficial explanations (See Table 3).

Table 3: Misclassified Responses in the Transformer-Based Completeness Classifier

Text	True Label	Predicted Label
The best possible outcome is that he is okay and the dog does not hurt him and no disease is spread to the rest of the colony. The worst	2	1

outcome is if the dog does attact him further hurting him and leaving him alone when he needs medical attention. My thinking was that there would be bad consequences for everyone if we broke protocol and then everyone in the colony ended up getting sick.		
The worst outcome to the situation is that something happens during that quarantine and there is a medical emergency. The best outcome is that there is no medical emergency and we can properly do the quarantine and then go to the medical bay for assessment. I thought about it this way because I figured if there was a medical emergency, we could cancel the quarantine and transport to the medical bay because lack of another option, but if not we could go with the best scenario and have the best outcome.	2	1
The best outcome is that Johnathan will be okay and the worst is that he will get more injured. I decided to keep Johnathan in the 24 hour quarantine, because there is rules that I have to follow as leader even though I have a personal relationship with Jonathan. If Jonathan is infected with something and he gets let out and gets more people sick it is my fault, and now more than one person is injured. If I make personal exceptions for the rules in place it will make other people try and break rules.	2	1
Trying to save Jonathan would put you and others at harm, but you would be able to save Jonathan's life. By choosing to save Jonathan, I would be able to say that I at least tried to save him rather than not even trying at all. I would rather say that I tried than looking the other way.	1	2
The worst case scenario in leaving the crew member in the airlock is the death of the crew member, potentially in a violent and painful way. The best case of opening the airlock would be the safety of the injured crew member, and the dog not having any sort of pathogen or other threat to the colony.	1	2

Perspectives Classifier

The development of a perspective classifier enables the investigation of how students position themselves in ethical reasoning. By automating this classification, key research questions can be explored, such as whether students who frame their responses in the first person engage more deeply with ethical dilemmas compared to those using other perspectives. Another consideration is whether perspective in ethical reasoning correlates with the actual decision made. Additionally, shifts in perspective use over time can be examined, particularly as students gain familiarity with the narrative or encounter increasingly complex dilemmas.

Interestingly, a substantial portion of students responded using second-person framing, writing responses in the form of direct instructions to a hypothetical decision-maker ("You should..."). This lead to a revision of the Perspective classification rubric to include this category alongside first-person ("I would...") and third-person ("He/she/they should...") justifications. To automate

classification, a transformer-based approach was developed, assigning responses to one of four categories:

First-Person (1) – Uses pronouns like I, me, my, we, our.

Second-Person (2) – Uses you, your, indicating an instructional or advisory tone.

Third-Person (3) – Uses he, she, they, their, framing reasoning in a detached, generalized manner.

Neutral (0) – No clear perspective detected/response too short.

Initially, the DeBERTaV3 transformer-based model was tested for perspective classification; however, it performed poorly, particularly in identifying second-person responses. The model consistently failed to recognize the advisory or instructional nature of "you" statements, likely because it was designed to classify text based on broader contextual patterns rather than explicit linguistic markers (see Table 4 and Figure 3).

Table 4: Transformer-Based Perspective Classifier Metrics:

Accuracy	Precision	Recall	F1 Score
0.8667	0.8093	0.8667	0.8286

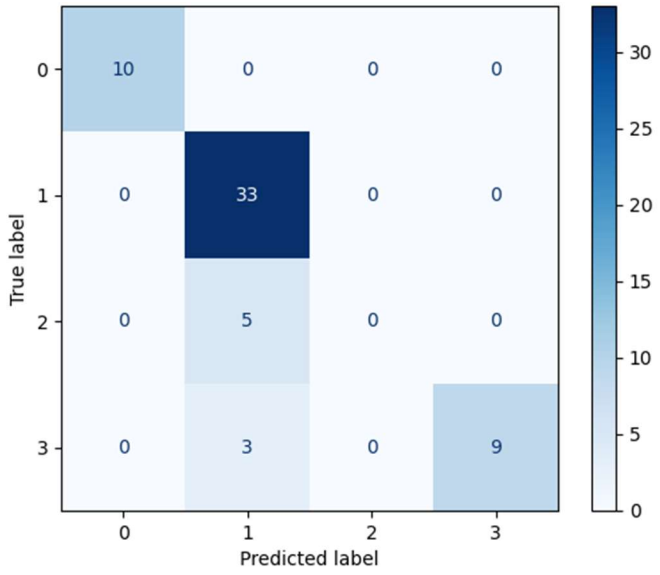

Figure 3: Confusion Matrix for Transformer-Based Perspective Classifier Predictions

Given the importance of reliably distinguishing perspectives in ethical reasoning, this limitation led to the development of the dictionary-based classifier, which explicitly relies on pronoun detection to assign perspective categories with greater accuracy. The classifier employed a hierarchical rule-based approach to determine the dominant perspective in student responses. First, responses were tokenized into individual words, with capitalized instances of "ME" excluded to prevent misclassification from unrelated contexts (e.g. Mechanical Engineering

abbreviation). The classifier then assigned a perspective category based on the presence of pronouns, following a structured priority system:

If a response contained second-person pronouns, it was classified as second-person (2) to reflect its instructional or advisory tone. If no second-person pronouns were found but first-person pronouns were present, the response was classified as first-person (1), indicating personal reflection. If neither second-person nor first-person pronouns appeared, the response was classified as third-person (3) by default, even if explicit third-person pronouns were not present, as their absence suggests an external, detached framing. However, if the response contained fewer than seven words, it was classified as neutral/unclassified (0) to account for cases where insufficient text prevented meaningful perspective identification.

We prioritized second-person responses to account for cases like "I think you should do XYZ," which, despite containing a first-person pronoun, primarily engage with the scenario through a second-person lens. Without this prioritization, such responses would have been classified as first-person, failing to capture the advisory or directive framing that distinguishes them from purely self-referential justifications. The classifier's performance was evaluated by comparing its predictions to manually labeled responses, using accuracy, precision, recall, and F1-score to assess reliability. The results indicated that this dictionary-based approach effectively identified and categorized student perspectives, enabling scalable analysis of ethical reasoning within *Mars!*'s decision-making framework (See Table 5 and Figure 4).

Table 5: Dictionary-Based Perspective Classifier Metrics:

Accuracy	Precision	Recall	F1 Score
0.9598	0.9767	0.9598	0.9643

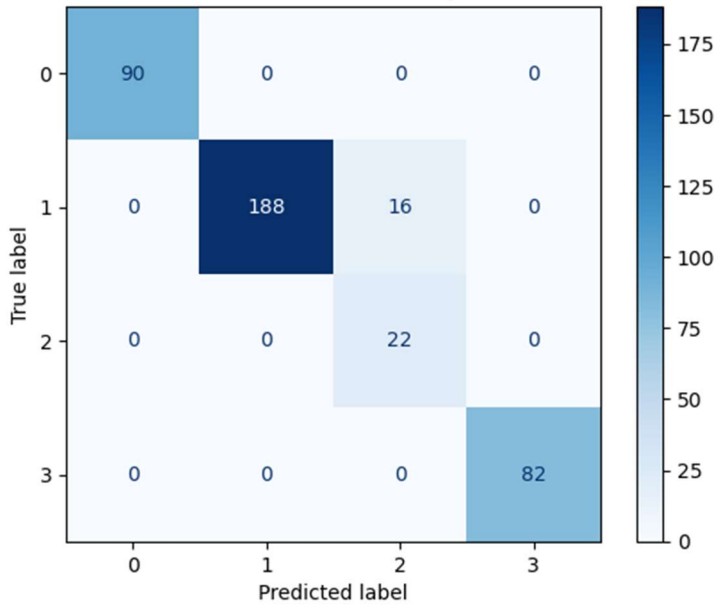

Figure 4: Confusion Matrix for Dictionary-Based Perspective Classifier Predictions

The dictionary-based classifier successfully distinguishes between first-person, second-person, and third-person perspectives, with high accuracy in clear-cut cases. Misclassifications occur in responses that mix perspectives (e.g., switching from "You should..." to "I believe..." within the same response). Continued refinement of the rules of the classifier may further enhance the metrics (See Table 6).

Table 6: Examples of Misclassified Responses in the Dictionary-Based Perspective Classifier

Text	True Label	Predicted Label
The worst outcome is that the husky has infected Jonathon and by bringing him inside the colony you are putting the colony at risk of being infected. The best outcome is that there is no infection and that you are able to rescue Jonathon from being injured any worse. I weighed the pros and cons of bringing Jonathon inside the colony but decided that Jonathon's life was important to me and that it would be wrong to leave him there with a vicious dog. The worst case definitely gave me pause and I considered leaving him out there but in the end I felt too much guilt leaving him.	1	2
The worst case scenario is the dog escapes and infects more of the colony. The best case is we only get Jonathon out and get him healed while the dog is still locked in the air chamber. I figured even if the dog did get out, me and the other people in the room would not allow the dog to escape. Also, Jonathon is a close colleague who just shared his desire to get back to his family, so if he died I would feel a huge sense of guilt. While I am the head of the Southern Hemisphere and let personal feelings get in the way, sometimes I feel it is necessary for those you care about.	1	2

Motives Classifier

The development of a motives classifier allows for the analysis of how students justify their ethical choices, particularly whether they prioritize personal benefit or broader social responsibility. Automating this classification enables the exploration of research questions such as whether students who emphasize social good are more inclined to justify bending rules for ethical reasons, while those focused on self-interest demonstrate a stronger adherence to established protocols. Another area of interest is whether the type of motive influences the depth of reasoning, with socially motivated justifications potentially leading to more complex ethical considerations. Additionally, tracking changes in motive use over time can provide insight into whether students shift toward more socially driven reasoning as they progress through the narrative. Responses in the Motives Classifier are categorized into three levels:

Neither (0) – The response does not clearly indicate a motive, lacking a discernible focus on either self-interest or social good.

Self-Interest (1) – The justification primarily emphasizes personal benefit, consequences, or advantages for the individual, with little regard for broader societal or communal impact.

Social Good (2) – The response prioritizes the well-being of others, collective responsibility, or ethical considerations that extend beyond personal gain.

The performance metrics for the motives classifier appear strong, with an F1 score of 0.9014, indicating high overall reliability in classification (see Table 7). However, an issue becomes evident upon examining the confusion matrix—the model fails to predict any self-interest (1) labels (see Figure 5). This suggests that while the classifier performs well in distinguishing between neither (0) and social good (2) responses, it struggles to recognize self-interest-driven justifications. Such a pattern often arises when categories are highly imbalanced, leading the model to favor the more frequent classes while effectively ignoring the underrepresented category.

Table 7: Transformer-Based Motives Classifier Metrics:

Accuracy	Precision	Recall	F1 Score
0.9333	0.8722	0.9333	0.9014

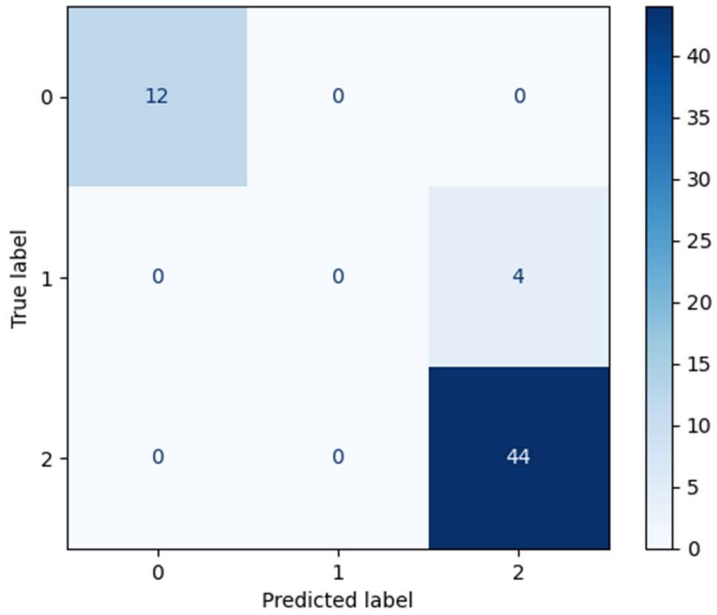

Figure 5: Confusion Matrix for Motives Classifier Predictions

Table 8 highlights misclassified responses in the motives classifier, illustrating the key challenge in training the model—the extreme imbalance of self-interest (1) responses. Given the low number of students who explicitly justified their decisions based on self-interest, responses that contained any element of self-interested thinking, even if they also included socially driven reasoning, were manually labeled as 1 to create a more balanced dataset. Despite this, the classifier still struggled to differentiate between self-interest (1) and social good (2) responses, often misclassifying responses that included both personal concerns and broader ethical

considerations as purely social good (2). This suggests that the model was unable to develop a strong distinction between mixed-motive responses and those that were primarily self-serving, likely due to the overwhelming dominance of social good responses in the dataset.

Table 8: Misclassified Responses in the Motives Classifier

Text	True Label	Predicted Label
The worst is that a disease could spread from Jonathan and the colony could be worried. The best is that **I can save Jonathan and make it known that I would be a leader that cares about the other members of the colony.**	1	2
The worst possible outcomes include getting myself sick or injured but in this situation I have to put my coworker before myself	1	2
Jonathan may be upset but its is important to do what is best for the colony. **This could have consequences for my relationship with jonathan.**	1	2
The worst case scenario is the dog escapes and infects more of the colony. The best case is we only get Jonathon out and get him healed while the dog is still locked in the air chamber. I figured even if the dog did get out, me and the other people in the room would not allow the dog to escape. Also, **Jonathon is a close colleague who just shared his desire to get back to his family, so if he died I would feel a huge sense of guilt. While I am the head of the Southern Hemisphere and let personal feelings get in the way, sometimes I feel it is necessary for those you care about.**	1	2

Reasoning Approach Classifier

The development of a Reasoning Approach Classifier enables the examination of how students apply ethical reasoning when justifying their decisions—whether they rely on strict rule adherence or take a more context-sensitive approach. Automating this classification allows for the investigation of research questions such as whether students who emphasize situated reasoning are more likely to consider exceptions and broader implications, while those who prioritize strict rule application consistently adhere to predefined policies. Another important consideration is whether reasoning approach correlates with the complexity of responses, with context-driven justifications potentially demonstrating more nuanced ethical reflection. Additionally, analyzing shifts in reasoning throughout the story may reveal whether students become more flexible in their ethical decision-making as they engage with increasingly complex dilemmas throughout the narrative. Responses in the Reasoning Approach Classifier are categorized into three levels:

Neither (0) – The response does not clearly indicate a reasoning approach, lacking a discernible focus on either strict rule adherence or context-driven decision-making.

Strict Rule Application (1) – The justification emphasizes following established rules, policies, or protocols without minimal consideration of situational factors or potential exceptions.

Situated Reasoning (2) – The response demonstrates flexibility, taking into account the specific context, nuances, and potential trade-offs involved in the ethical dilemma.

The performance metrics for the reasoning approach classifier are worse than the motives classifier, with an F1 score of 0.7581, indicating moderate reliability in classification (see Table 9). However, similarly to the motives classifier, upon examining the confusion matrix, a critical issue emerges—the model fails to predict any strict rule application (1) labels (see Figure 6). This suggests that while the classifier effectively differentiates between neither (0) and situated reasoning (2) responses, it struggles to recognize justifications that emphasize strict adherence to rules and protocols. As with the motives classifier, this pattern is likely due to an imbalance in the dataset, where responses favoring strict rule application are underrepresented. As a result, the model defaults to predicting the more frequent categories, overlooking instances where students rigidly follow established guidelines.

Table 9: Transformer-Based Reasoning Approach Classifier Metrics:

Accuracy	Precision	Recall	F1 Score
0.8333	0.6962	0.8333	0.7581

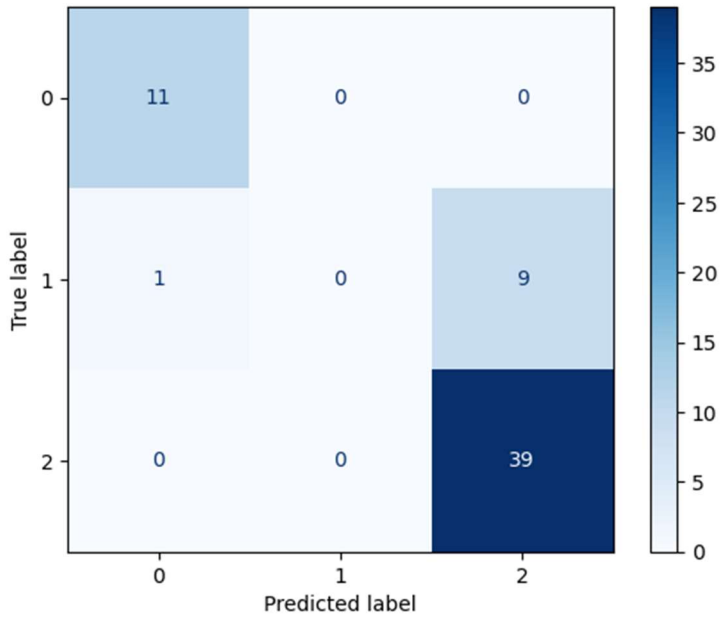

Figure 6: Confusion Matrix for Reasoning Approach Classifier Predictions

Table 10 highlights misclassified responses in the reasoning approach classifier, showing the limitation in the model's ability to distinguish strict rule application (1) from situated reasoning (2). Due to the low number of responses that strictly adhered to rules without considering context, any justifications that emphasized policy enforcement, protocol adherence, or rule-

following were manually labeled as 1 in an attempt to balance the dataset. However, as shown in the table, the classifier frequently misclassified these responses as situated reasoning (2), likely because many students justified rule-following by also referencing broader ethical considerations. This suggests that the model struggled to recognize responses that framed rules as absolute versus those that invoked them within a larger ethical framework.

Table 10: Examples of Misclassified Responses in the Reasoning Approach Classifier

Text	True Label	Predicted Label
The worst possible outcome of my decision is that Johnathan dies. He could bleed to death or he could succumb to a virus the dog was carrying. The best possible outcome is that he lives, and you are able to cure him. While this decision could kill Johnathan, it i very important not to put the entire team in jepardy. Johnathan chose to go after the alien dog, and this is the result of his rash decision. **The rules in this and in any colony are to protect the masses and not risk their safety for the minority.** Johnathan could have a lethal illness that would spread to the entire colony if you let him out early and you cannot let that happen.	1	2
The worst is he dies. My thinking is that he did this to himself without the best interests of the colony in his mind. **As the leader it is important to always have the wellbeing of your colony or group over any other beings.** In this case he chose the dog over the wellbeing of his comrades and suffered the consuquence. The best is that the dog lets go and that jonathan has no infection or disease. However, whatever the outcome is he put the colony second so why would we save him.	1	2
The best outcome is that Johnathan will be okay and the worst is that he will get more injured. **I decided to keep Johnathan in the 24 hour quarantine, because there is rules that I have to follow as leader even though I have a personal relationship with Jonathan.** If Jonathan is infected with something and he gets let out and gets more people sick it is my fault, and now more than one person is injured. If I make personal exceptions for the rules in place it will make other people try and break rules.	1	2
Best outcome: After 24 hours Johnathon is still okay and the dog left him alone and we can help his wounds. Worst outcome: Johnathon dies or is infected. **While it could end with Johnathon dying, the quarentine rule is set in place for a reason, and we must follow it.** It can benefit the overall community if Johnathon were to end up infected. This way the colony remains safe. Although Johnathon is a dear friend, the safety of the colony comes first.	1	2
You have to follow the protocol especially in such a curious instance.	1	0

Discussion

The implementation of NLP-based classifiers for ethical reasoning in *Mars! An Ethical Expedition* has demonstrated strong initial effectiveness, particularly in distinguishing levels of completeness and identifying perspective framing. The completeness classifier performed well, achieving an F1-score of 0.9162, suggesting that it can effectively differentiate between irrelevant, partial, and fully developed justifications. While this level of performance indicates that automated methods can provide a broad understanding of student engagement into student engagement with ethical dilemmas—potentially aiding instructors in assessing classroom-wide trends to support ABET accreditation—further refinement would be necessary for reliable individual-level grading. The perspective classifier also performed well after transitioning from a transformer-based approach to a dictionary-based model, which enabled more precise classification of responses into first-person, second-person, and third-person categories.

One of the most unexpected findings was the prevalence of second-person framing in student responses. Many students articulated their reasoning in the form of directives to a hypothetical decision-maker ("You should do X"), rather than situating the justification in their own perspective ("I would do X") or in a third-person analysis ("Jonathan should do X"). This discovery prompted a revision of the classification rubric to formally include second-person responses as a distinct category. The presence of second-person reasoning suggests that students may be treating the dilemmas as if they were narrating or guiding a character rather than reflecting solely on their own ethical stance. Further investigation is needed to determine whether this framing impacts the depth or quality of ethical reasoning, particularly in comparison to first-person or third-person responses.

Beyond perspective, the study revealed consistent trends in ethical justification, with students overwhelmingly favoring social good over self-interest and situated reasoning over strict rule application. Very few responses prioritized self-interest as a driving motive, suggesting that students generally framed their ethical responsibilities in terms of collective well-being rather than personal gain. Additionally, most students engaged in context-sensitive decision-making rather than rigid rule adherence, demonstrating an awareness of how ethical choices must be adapted to situational nuances. This pattern suggests that engineering students may naturally approach ethical reasoning as a balancing act, weighing competing factors rather than strictly applying predefined ethical principles. The relative lack of strict rule-based justifications raises important questions about how ethical principles are traditionally taught in engineering education—while ethical frameworks and codes of conduct provide essential guidance, the results suggest that students may internalize ethical decision-making as a flexible, context-driven process rather than a rigid application of rules.

Challenges & Limitations

While the use of NLP models for ethical reasoning classification in *Mars!* has demonstrated promising results, several challenges and limitations must be considered. These include data imbalance in student responses, potential bias in automated classification, and ethical concerns

surrounding the use of AI-driven assessment tools. Addressing these limitations is essential for refining the accuracy and reliability of NLP-based approaches in ethics education.

One of the primary challenges was data imbalance, particularly in the distribution of motive and reasoning approach classifications. Student responses were overwhelmingly skewed toward social good for motive and context-driven reasoning for reasoning approach, with far fewer responses classified as self-interest or strict rule application. This imbalance made training machine learning models difficult, as the models had fewer examples of certain reasoning styles, reducing classification accuracy for underrepresented categories. This imbalance affected different components of the NLP pipeline in distinct ways. The dictionary-based classifier for perspective was less affected, as pronoun detection remained reliable across response types. However, the DeBERTaV3 model for motive and reasoning approach struggled to generalize for less frequent reasoning styles, leading to lower performance metrics when classifying responses that prioritized personal benefit or rigid rule adherence. The completion scoring model had a more balanced distribution of responses in each category and performed much better than the other transformer-based classifiers. Future iterations of the models will incorporate additional data collection and expanded training datasets to improve classification accuracy across all categories. To address class imbalances, techniques such as data augmentation and synthetic response generation may be implemented to ensure more balanced learning. Further refinements will also include testing alternative transformer architectures and fine-tuning hyperparameters to enhance model robustness. Additionally, manual review of misclassified and low-frequency response types will be conducted to improve label consistency and ensure fair and accurate classification.

Automated classification introduces the risk of bias, both in the underlying models and in the annotation process used to train them. One major concern is training data bias, as the classifier is trained on student-generated responses, meaning any biases present in the original dataset may be reinforced by the models. If certain ethical reasoning styles are more common among the sample population, the models may over-prioritize those styles, leading to underrepresentation of alternative perspectives. Additionally, cultural and linguistic biases may affect classification accuracy. Students from different cultural backgrounds may frame ethical reasoning differently, using distinct rhetorical structures, justifications, or implicit reasoning patterns that the model may not recognize if it was primarily trained on responses from a homogeneous sample. Similarly, English proficiency levels among students, particularly for English language learners, could impact response length, complexity, and phrasing, potentially influencing classification outcomes. If the training dataset lacks sufficient diversity in linguistic styles and cultural perspectives, the model may generalize poorly to broader student populations. Future work will examine the dataset's representativeness and explore techniques to improve classification fairness and ensure more equitable assessment across diverse student cohorts.

Bias is also a concern in the DeBERTaV3 classifier for completeness, motive, and reasoning approach, as it is trained on a dataset labeled by human annotators, meaning that subjective judgments during annotation may introduce unintended biases. To mitigate this, inter-rater reliability measures will be implemented, ensuring that multiple annotators consistently apply the

same criteria when labeling responses. By comparing results across annotators and resolving discrepancies before model training, we can improve the consistency and objectivity of the labeled data, reducing potential bias in the final classification models. Ongoing validation through expert review will be conducted to ensure that automated classifications align with human assessments.

While NLP-based assessment offers scalability and efficiency, the use of automated classifiers to evaluate ethical reasoning presents challenges related to nuance and contextual interpretation. Ethical decision-making is inherently complex and reducing it to algorithmic classification risks overlooking the depth and intent behind student justifications. Currently, the completeness classifier assesses individual responses but is more informative when aggregated at the classroom level. With an accuracy of 91.67%, the model performs well in distinguishing between irrelevant, partial, and complete responses. However, at the individual level, this means that approximately 1 in every 12 student responses may be incorrectly classified—a margin of error that is too high for high-stakes individual assessment. While this level of accuracy is insufficient for grading individual students, it is still valuable for classroom-wide analysis. For example, if the classifier predicts that 90% of responses in a class are partial or complete, the 91.67% accuracy suggests that this estimate is fairly reliable. While individual misclassifications occur, the aggregate trend provides strong evidence of student engagement with ethical dilemmas, which can support ABET accreditation efforts. Future improvements—such as expanding the training dataset, refining classification thresholds, and incorporating additional linguistic features—may improve accuracy to the point where individual assessment is more feasible. Additionally, if an in-game implementation can accurately identify incomplete responses in real time, it could prompt students to expand their reasoning before submission, enhancing engagement and reflection on ethical decision-making. Future research will explore ways to integrate AI-assisted feedback with instructor guidance and peer discussions, ensuring that automated assessment remains pedagogically valuable and aligned with the complexities of ethical reasoning.

Conclusion

This study demonstrates the potential of NLP-based classification for analyzing ethical reasoning in *Mars! An Ethical Expedition*, providing a scalable approach to assessing student justifications in a narrative-driven ethics intervention. The classifiers successfully categorized responses based on completeness and perspective, but failed to reliably categorize responses based on motives and reasoning approach. Notably, the overwhelming preference for social good over self-interest and for situated reasoning over strict rule adherence suggests that ethical decision-making in context is more nuanced than rigid application of ethical principles. Additional refinements to all models—such as expanded training data, improved model fine-tuning, and enhanced handling of imbalanced categories—are necessary before these models can be reliably used for individual assessment. Future research will focus on improving classification balance, incorporating inter-rater reliability measures to ensure consistency in labeled data, and refining models for real-time in-game analysis. Ultimately, this work contributes to the broader effort of integrating AI-driven

assessment tools into engineering ethics education, supporting both large-scale evaluation and deeper classroom discussions on ethical decision-making.

References

1. J. R. Rest, D. Narvaez, S. J. Thoma, and M. J. Bebeau, "DIT2: Devising and testing a revised instrument of moral judgment," *J. Educ. Psychol.*, vol. 91, no. 4, pp. 644–659, Dec. 1999.
2. Q. Zhu, C. B. Zoltowski, M. K. Feister, P. M. Buzzanell, W. C. Oakes, and A. D. Mead, "The development of an instrument for assessing individual ethical decision making in project-based design teams: Integrating quantitative and qualitative methods," in *Proc. 2014 ASEE Annu. Conf. Expo.*, Indianapolis, IN, USA, Jun. 2014, pp. 24–1197.
3. L. Kohlberg and R. H. Hersh, "Moral development: A review of the theory," *Theory Pract.*, vol. 16, no. 2, pp. 53–59, Apr. 1977.
4. N. E. Canney and A. R. Bielefeldt, "Validity and reliability evidence of the engineering professional responsibility assessment tool," *J. Eng. Educ.*, vol. 105, no. 3, pp. 452–477, Jul. 2016.
5. J. Borenstein, M. Drake, R. Kirkman, and J. Swann, "The test of ethical sensitivity in science and engineering (TESSE): A discipline-specific assessment tool for awareness of ethical issues," in *Proc. 2008 ASEE Annu. Conf. Expo.*, Pittsburgh, PA, USA, Jun. 2008, pp. 13–1270.
6. T. N. Wagner, D. D. Burkey, R. T. Cimino, S. Streiner, K. D. Dahm, and J. Pascal, "Collective vs. Individual Decision-Making in an Engineering Ethics Narrative Game," in *Proc. ASEE Annu. Conf. Expo.*, Portland, OR, USA, Jun. 2024. doi: 10.18260/1-2—48470.
7. "Criteria for accrediting engineering Programs, 2025 - 2026 - ABET," *ABET*, Feb. 19, 2025. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2025-2026/
8. W. X. Zhao et al., "A survey of large language models," *arXiv*, Jan. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.18223
9. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Representations (ICLR)*, Scottsdale, AZ, USA, May 2013.
10. Z. Chu et al., "History, development, and principles of large language models—An introductory survey," *arXiv*, Feb. 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.06853
11. J. S. Brown, A. Collins, and P. Duguid, "Situated cognition and the culture of learning," *Educ. Res.*, vol. 18, no. 1, pp. 32–42, Feb. 1989.
12. J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*, Cambridge, U.K.: Cambridge Univ. Press, Sep. 1991.
13. H. L. Dreyfus and S. E. Dreyfus, "The ethical implications of the five-stage skill-acquisition model," *Bull. Sci. Technol. Soc.*, vol. 24, no. 3, pp. 251–264, Jun. 2004.
14. J. P. Gee, *What Video Games Have to Teach Us About Learning and Literacy*, New York, NY, USA: Palgrave Macmillan, 2003.
15. V. J. Shute and F. Ke, "Games, learning, and assessment," in *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, New York, NY, USA: Springer, May 2012, pp. 43–58.

16. V. J. Shute, “Stealth assessment in computer-based games to support learning,” in *Computer Games and Instruction*, S. Tobias and J. D. Fletcher, Eds., Charlotte, NC, USA: Information Age Publishers, 2011, pp. 503–524.
17. R. J. Mislevy and G. D. Haertel, “Implications of evidence-centered design for educational testing,” *Educ. Meas. Issues Pract.*, vol. 25, no. 4, pp. 6–20, Dec. 2006.
18. P. He, J. Gao, and W. Chen, “DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing,” *arXiv (Cornell University)*, Jan. 2021, doi: 10.48550/arxiv.2111.09543.