The Pennsylvania State University

The J. Jeffrey and Ann Marie Fox Graduate School

ADAPTATION, APPLICATION AND ANALYSIS OF A CONTENT ANALYSIS TOOL FOR ESSAY FEEDBACK IN MIDDLE SCHOOL SCIENCE CLASSROOMS

A Thesis in

Computer Science and Engineering

by

Mahsa Sheikhi Karizaki

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2024

The thesis of Mahsa Sheikhi Karizaki was reviewed and approved by the following:

Rebecca J. Passonneau
Professor of Computer Science and Engineering
Thesis Advisor

Wenpeng Yin
Assistant Professor of Computer Science and Engineering

ChanMin Kim
Associate Professor of Education

Chitaranjan Das
Distinguished Professor of Computer Science and Engineering
Program Head

ABSTRACT

This thesis investigates methods to apply a pre-existing software tool, PyrEval, in middle school classrooms for formative assessment of science essays, analyzes characteristics of writing that affect its performance, and compares PyrEval with two other approaches. These are a recently developed end-to-end neural network for assessment of lab reports, and large language models (LLMs). PyrEval has been under continuous development over the past decade. In recent years, it has been applied to formative assessment of science explanation essays written by middle school students in Wisconsin public schools. PyrEval performs well in identifying whether an essay expresses important target ideas from a curriculum where students learn about energy, mass and speed through simulated roller coaster experiments. As a result, PyrEval supports feedback to students and teachers, and this feedback has been shown to lead to improved understanding of science concepts in students' revised essays. This thesis reports our method to align PyrEval with essay rubrics, comparison of multiple semantic vector methods for use in PyrEval, in-depth analyses of several aspects of student writing quality and the impact on PyrEval performance, and finally, comparison of PyrEval with two other automated assessment methods.

Previous work by team members compared two vector dictionary methods, meaning methods that compute a fixed vector for each word string, and found that a matrix factorization method for words and phrases yielded the best performance. My work investigates a variety of methods to assess whether performance can be improved. These methods include modification of the training corpus to be more specific to the student writing, modification of the vector space for the fixed word and phrase vectors, and use of contextualized vectors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have supported me throughout my time in the graduate program and in the completion of my thesis. I am particularly grateful for the opportunity to be part of the Natural Language Processing Lab at Penn State and to work under the guidance of my advisor, Professor Rebecca J. Passonneau. Her mentorship and advice have been invaluable not only for this thesis but for my overall academic and personal growth. I would also like to express my appreciation to Professors ChanMin Kim and Wenpeng Yin for their participation in my thesis committee and for their valuable feedback. My gratitude extends to Yanjun Gao and Purushartha Singh, whose work with PyrEval I have continued throughout my thesis.

In addition, I acknowledge that this work was partly supported under the NSF DRK award 2010351. The findings and conclusions represented in this work do not necessarily reflect the view or the policy of the Government, and no official endorsement should be inferred.

Chapter 1
Introduction

An important aspect of science writing is explanation, and science writing has been found to help students learn[3]. Alongside the benefits of science writing, there are high costs to teachers for designing the assignments, and providing ongoing feedback to students to help them revise. This thesis is an examination of several aspects of anM assessment tool based on NLP and other AI techniques applied to middle school science instruction to help students articulate and comprehend scientific concepts. AI is becoming more widely used in all aspects of human activity, including education. For education, there are aspects of AI that could be problematic, such as whether automated feedback might be misleading or less accurate than human feedback, as well as whether automated assessment could be equitable when applied to diverse groups of students. This thesis describes how an existing content analysis tool was customized for a web delivered middle school science curriculum in which students write essays, and revise them based on automated feedback. Further, we compare different semantic vector methods for use in PyrEval to identify which perform best. Finally, we analyze two key aspects of performance: what are the factors that affect the accuracy of the feedback and does it assess students fairly. During this project over 2000 essays were collected over a three-year deployment in Madison, WI public schools.

The thesis chapters are organized as follows. Chapter 2 discusses related work in the field. In Chapter 3, we discuss the datasets used in this project, detailing the sources, and the collection and annotation of the data.

Chapter 4 focuses on how PyrEval functions and its alignment with a priori rubrics. This chapter details how PyrEval was adapted to bridge the gap between automated essay assessment and the use of analytic rubrics. Here we specifically focus on two rubrics designed to evaluate middle school essays in which students explain their roller coaster designs, emphasizing energy concepts.

In Chapter 5, we compare different types of vector representation for PyrEval to determine which performs best on the ground truth dataset. Chapter 6 analyzes student writing and idea distinctiveness, using distributions of cosine similarities to reveal differences in the distinctiveness of main ideas and the clarity of students' statements. In Chapter 7, we compare PyrEval to another method called VerAs, which is an end-to-end neural architecture. This comparison evaluates the data requirements and accuracy of PyrEval relative to VerAs, providing insights into their respective strengths and weaknesses in automated content assessment.

The conclusion summarizes our contributions, including identification of PyrEval's strengths and weaknesses for formative feedback on science explanations.

I have four publications, one as first author, that draw upon work presented in this thesis. My first publication, "How Well Can You Articulate that Idea? Insights from Automated Formative Assessment" [16], where I am the lead author, is discussed in chapter 6. My second publication focuses on linguistic discrimination in an automated assessment tool and is under review. My third publication, "VerAs: Verify then Assess STEM Lab Reports" [17], is detailed in chapter 7. My fourth publication examines essay quality change when revised[18].

Chapter 2
Related Work

In the area of educational technology, particularly within STEM subjects, there's a growing body of research focused on the development and application of automated tools designed to enhance formative feedback for students solving open-ended questions. Formative feedback, which is provided during the progression of a unit or course to aid in further learning, proves to be most effective when it addresses the specifics of a problem—namely the what, how, and why—rather than merely confirming the correctness of answers[1].

Studies conducted by Linn and her team have explored the potential of automated guidance to support students, especially at the middle school level, in refining their short-answer responses. These investigations have explored different aspects of feedback provision: comparing the efficacy of automated feedback both on its own and when supplemented with insights into the feedback's personalized aspect[4], alone or in combination with students providing feedback on a sample essay[5], and alone or in paired with an interface that simulate the revision process[6].

The common tool across these studies, the C-rater-ML[7], demonstrated a notable correlation (0.72 Pearson correlation) with human-assessed feedback[4], underlining its utility in encouraging more effective use of evidence in scientific explanations—a crucial skill given students' typical struggles with revising their work. The combination of automated feedback with clear examples of how to revise was identified as the most advantageous approach.

Furthermore, the Concord Consortium's research[8,9,10], mainly involving high school students, aimed at deepening students' comprehension of uncertainty within scientific contexts[11]. Utilizing the C-rater-ML tool, these studies achieved Quadratic Weighted Kappa (QWK) scores ranging from 0.78 to 0.93 in alignment with human ratings, depending on the specific study. The first study found that students overwhelmingly improved their revisions given the automated feedback[10]. The second study compared generic argumentation feedback to student-specific feedback through use of C-rater-ML, with the latter leading to greater improvements in revisions[9]. The third study compared feedback on argumentation writing alone or in combination with feedback on students' use of the science simulations and resulting data[8]. Revisions from students in both conditions improved over the original responses; only those who also received feedback on the simulations improved their use of data.

Distinct from the previously mentioned research on short answer responses, our investigation extends the application of formative feedback to science explanation essays, which require students to elucidate multiple concepts. There is a notable scarcity of research on automated feedback systems for essay revisions. Zhang et al.[12] introduced eRevise, an innovative tool offering rubric-based feedback aimed at enhancing students' evidence usage in source-based essays. Through comparative analysis, this tool emerged as particularly effective, owing largely to its use of word embeddings to furnish tailored feedback. Testing with middle school students confirmed that eRevise significantly boosted the quality of essay revisions.

Our research employs PyrEval[2], an open-source software using word embeddings to evaluate essays against a rubric detailing essential ideas for student explanations[13]. Preliminary findings suggest that PyrEval is adept at recognizing the presence or absence of key ideas within essays, although its accuracy varies based on the uniqueness of the idea in question and the clarity of the students' writing (see chapter 5).

Conversations with educators and learners regarding instances where PyrEval identified an idea as missing—despite students' belief that they had included it—indicate that even partially inaccurate feedback can provoke thoughtful reflection among students about their expression of ideas. This underscores the benefits of automated feedback, not only in directly improving student revisions but also in fostering deeper engagement with the material and enhancing critical thinking skills.

Chapter 3

Creation and Use of Ground Truth Datasets

Creating an accurate ground truth is a crucial step in the development and evaluation of an automatic assessment tool like PyrEval. Our methodology involved a process where 39 student essays were selected, and three evaluators manually annotated each essay. The focus was on identifying whether the students stated the main ideas specified in the curriculum, and if so, pinpointing the exact sentences where these articulations occurred. Through a collaborative process, the evaluators discussed each essay's assessment until a consensus was reached, ensuring a robust ground truth for subsequent validation.

Collection and Annotation of Post-Engagement Essays

An additional collection of 120 essays was gathered from the students following their engagement with PyrEval. This compilation comprised two categories: 60 original essays and 60 revised essays. The latter was obtained after feedback was provided by PyrEval, wherein students were requested to revise their essays based on the PyrEval feedback. Following that, the dataset underwent inter-annotator agreement by a pair of evaluators.

Creation of MidPhys Dataset

Complementing our evaluation of PyrEval, we assembled another essential dataset, MidPhys, derived from assessment questions and student responses from a decade of historical data. This dataset serves two purposes: it is a rich source for data mining, and it provides an extensive corpus for fine-tuning large language models. By leveraging this dataset, we can refine

PyrEval's vector representation, with the aim of enhancing its accuracy in educational assessments of middle school physics essays.

Chapter 4
Alignment of PyrEval Models to an a Priori Rubrics

PyrEval was originally developed to be a fully automated method to assess the importance of ideas in short passages written to the same prompt. In applying PyrEval to an existing curriculum, where students receive an essay prompt that elicits a pre-defined set of important ideas, spelled out in a rubric, it is necessary to adapt PyrEval so that the content model is aligned to the rubric. In the adaptation of the PyrEval system addressed in this chapter, we aim to bridge the gap between automated essay assessment and the application of an analytic rubric. There are two rubrics designed for evaluating middle school essays where students explain their roller coaster designs with reference to energy concepts, such as the relation between potential and kinetic energy.

Early in the project to apply PyrEval to the roller coaster curriculum, we had no existing essays from which to develop a content model. At that time, we mined phrases from a historical dataset of middle school physics essays provided by our collaborators. These essays were not from the same curriculum, or written to the same prompt, so we mined relevant phrases that corresponded to the elements of the rubric[13]. In subsequent years, we could curate a content model created by PyrEval from high quality essays for the current curriculum.

A PyrEval content model is called a pyramid, which is essentially a list of weighted Content Units (CUs). A CU consists of phrases extracted from different reference passages that state the same idea. Given ***n*** reference passages, a CU will contain at most n phrases stating the same idea, each from a different reference passage; previous work has shown that 5 reference passages is sufficient. Therefore, automatic creation of a pyramid will lead to the establishment of a range of weights for Content Units (CUs), with significant ideas assigned a weight of 5 to less critical ideas with weights from 4 to 1. PyrEval then utilizes this weighted pyramid model to assess content by uniquely matching ideas in a student's passage to the CUs. Here we aim for a pyramid where there are exactly the number of weight 5 pyramids as there are ideas in the essay rubric. In brief, from 5 high quality essays written to a given essay prompt, we automatically create a pyramid, then curate it, meaning we modify it manually, to arrive at the one-to-one alignment of weight 5 CUs to the rubric.

The implementation of PyrEval involves a three-stage pipeline[2] as depicted in Figure1. The initial stage focuses on pre-processing, where complex sentences are broken down into clauses using a rule-based parser, and these clauses are then transformed into semantic vectors. In the subsequent stage, the EDUA algorithm, a specialized set partition algorithm, is employed to group similar vectors from the reference passages into CUs. The final stage of the pipeline leverages the WMIN algorithm, a greedy maximal independent set algorithm, to match students' ideas to the appropriate CUs.

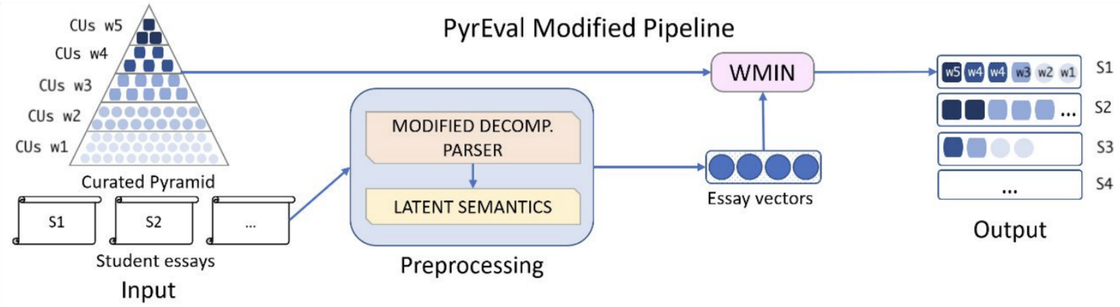


Figure1: PyrEval PipeLine

Given that PyrEval operates on a greedy search algorithm, the sequencing of the CUs is critical for the final accuracy of the assessment. Changing the order of the CUs in the pyramid, changes the order in which WMIN tries to match student phrases to the model, and therefore changes the assessment. We tested our curated pyramids on a ground truth dataset (see Chapter 3 above) . To retain the order of CUs produced by EDUA, and then preserved in our curated pyramid, while still referencing the CUs by the indexing used in the rubric, we introduced a dictionary in the MongoDB database. The dictionary maps the top CUs—representing the main ideas from the rubric—in an order that is both easy to understand and coherent to teachers and students.

Chapter 5
Comparing Multiple Vector Methods for PyrEval

As noted above in Chapters 1 and 3, PyrEval is an off-the-shelf tool that has a public GitHub repository and has been used in previous work on student essays[13]. PyrEval is highly modular, allowing the user to choose among different vector representations for words and phrases, and it has hyper-parameters that the user can experiment with to choose values that achieve the highest accuracy.

Here we compare six types of vector representation for PyrEval to choose the one that performs best on the ground truth dataset described in the previous section. Our previous work[13] found that WTMF vector representation, which relies on a pretrained vector dictionary using a matrix factorization method, outperformed GloVe word vectors. Early work tested contextualized representations, but here we aimed to evaluate the impact on WTMF of corpus modification and vector space adjustment or using contextualized PLM vectors on PyrEval accuracy.

WTMF using the WTMF Corpus (Baseline)

In our first experiment, we performed grid search to find the optimal values for the two hyperparameters for the original WTMF vector dictionary. This dictionary is created from a corpus distributed with the original WTMF code, consisting of definitional sentences from three electronic lexicons, plus the Brown corpus[15]. It has 393,666 sentences, over 4M words, and a lexicon of nearly 100K words. As discussed in our previous work[13], a lexicon extracted from this corpus and a fixed vector dictionary trained with WTMF is larger and of higher quality than a lexicon extracted from and trained on a Gigaword subset 4.5 times the size of the WTMF corpus; further, the vector dictionaries trained with WTMF on these corpora perform better than using GloVe (Pennington et al., 2014). The first experiment performs a new a grid search on PyrEval, given that the parameters last optimized were based on a historical dataset[13]. Considering the curriculum and students have since changed, there was a need to re-evaluate the model's performance under different parameters to have a baseline for comparison.

While accuracy was high, we found that certain ideas in the curriculum seemed to be less differentiated from one another than others, such as main idea 2 “potential energy and kinetic energy are inversely related” and main idea 3 “potential energy and kinetic energy add up to the same total energy at any point on the track”.

WTMF Pipeline Using a Mixed Corpus as Input

Our second experiment was motivated by the hypothesis that training WTMF on a corpus that included middle school physics essays might improve performance, given that middle school writing has many characteristics that differentiate it from the original corpus WTMF was trained on. Also, we hypothesized that adding more text about middle school physics might help WTMF differentiate some of the curriculum ideas. We created a mixed corpus by concatenating the Weiwei corpus, initially used with the original WTMF pipeline, with a middle school physics corpus as described in chapter 3. The WTMF pipeline was run on this new corpus to generate a new vector dictionary, which was subsequently input into PyrEval for another grid search experiment.

Word Embedding Refinement

The third experiment was inspired by Yu's paper, "Refining Word Embeddings for Sentiment Analysis". This paper introduces a methodology for refining word embeddings by adjusting vector representations. The objective is to bring semantically and sentimentally similar words closer in the vector space while distancing them from sentimentally dissimilar words. In our experiment, we deviated from the original sentiment-based ranking to employ a different ranking mechanism.

We employed Term Frequency-Inverse Document Frequency (TF-IDF) as our ranking criterion for the lexicon. The decision was based on two primary attributes:

- The degree to which a word is characteristic of the physics domain of interest.
- The specificity or distinctiveness of the word within that domain.

For example, terms like "energy," "potential," "kinetic," and "mass" are highly characteristic of the physics domain and would thus receive high ranks based on the first criterion. However, when it comes to specificity, "potential" and "kinetic" would receive higher ranks than "mass" or "energy."

By calculating the TF-IDF scores, we were able to automatically simulate the two criteria for all words in the middle school physics corpus, treating each sentence as a document. A high TF-IDF score in our ranking system indicates that a word is more distinctive.

After that, the refined word embeddings were incorporated into PyrEval, followed by the execution of an additional grid search. This step aimed to assess the impact of the newly adjusted embeddings on the model's performance.

Contextualized Embeddings

Our fourth through sixth experiments marked a shift in our methodology by integrating BERT (Bidirectional Encoder Representations from Transformers) into PyrEval and using max pooling for generating vector representations. This approach was divided into three distinct parts: employing the original BERT model, fine-tuning BERT on the Middle School Physics corpus, and a novel experiment where BERT's output was concatenated with that of the WTMF dictionary. This approach was designed to explore the efficacy of contextualized embeddings in enhancing the representational accuracy of the students' essays and pyramid segments within PyrEval. The final results are summarized in Table1 below.

Experiment Number	Edge Threshold	Top K SCUs	Accuracy
WTMF Using WeiWei's corpus (Baseline)	0.55	3	0.79
WTMF Pipeline Using a Mixed Corpus as Input	-0.1	3	0.68
Word Embedding Refinement	-0.1	3	0.71
BERT	0.85	3	0.75
BERT Fine-tuned	0.83	3	0.75

Table1: Comparison of Six Semantic Vector Methods

PyrEval Performance on Ground Truth

As shown in the preceding section, for our current project assessing middle school students' essays on the physics of roller coasters, we found that the original WTMF vector dictionary trained on the definitional corpus provided by the developers of WTMF[14] performed best on our ground truth data. In this section, we analyze PyrEval accuracy in the feedback provided to students. The feedback is in the form of a checklist for each of six main ideas that the essay prompt is designed to elicit. Here, both positive accuracy (sensitivity; the idea is present) and negative accuracy (specificity; the idea is not present) are important.

In assessing PyrEval's performance, we employed a nuanced approach by measuring positive accuracy (sensitivity)—instances where PyrEval correctly identified the expression of main ideas in a student's essay—and negative accuracy (specificity), which measures PyrEval's ability to recognize accurately when a main idea was not stated as shown in Table2. Additionally, we calculated the tool's overall accuracy in evaluating student essays. This comprehensive accuracy assessment allowed us to understand PyrEval's efficacy in not only recognizing the presence of main ideas but also its sensitivity to their absence. Both are important for providing students with accurate feedback for how to improve in a revised version of their essay.

Dataset	Positive Accuracy	Negative Accuracy	Total Accuracy
39 Ground Truth Essays	80.64%	76.56%	79.05%
60 Original Essays	73.73%	77.14%	74.72%
60 Revised Essays	77%	55.32%	74.17%

20 O+R Essays	75.53%	70.39%	74.44%
All Essays Combined	76.88%	70.05%	75.47%

Table2: PyrEval positive, negative and total accuracies (as percentages)

Accuracy by Main Idea

Beyond the positive, negative, and overall classification of accuracy, we explored a more detailed analysis by examining PyrEval's performance relative to each main idea within the curriculum that the essay prompt reminds students to include, as you can see in Table3. Our findings suggest that PyrEval's accuracy is influenced by the distinctiveness of an idea compared to others within the curriculum, and whether it is a definitional idea or not. Main ideas that were more unique and well-defined were easier for PyrEval to detect and evaluate correctly.

Dataset	Main Idea 1	Main Idea 2	Main Idea 3	Main Idea 4	Main Idea 5	Main Idea 6
39 Ground Truth Essays	76.92%	82.05%	69.23%	89.74%	71.79%	84.62%
60 Original Essays	63.33%	56.66%	66.66%	91.66%	86.66%	83.33%
60 Revised Essays	63.33%	61.66%	76.66%	86.66%	86.66%	70%
20 O+R Essays	63.33%	59.16%	71.66%	89.16%	86.66%	76.66%
All Essays Combined	66.66%	64.77%	71.06%	89.3%	83.01%	78.61%

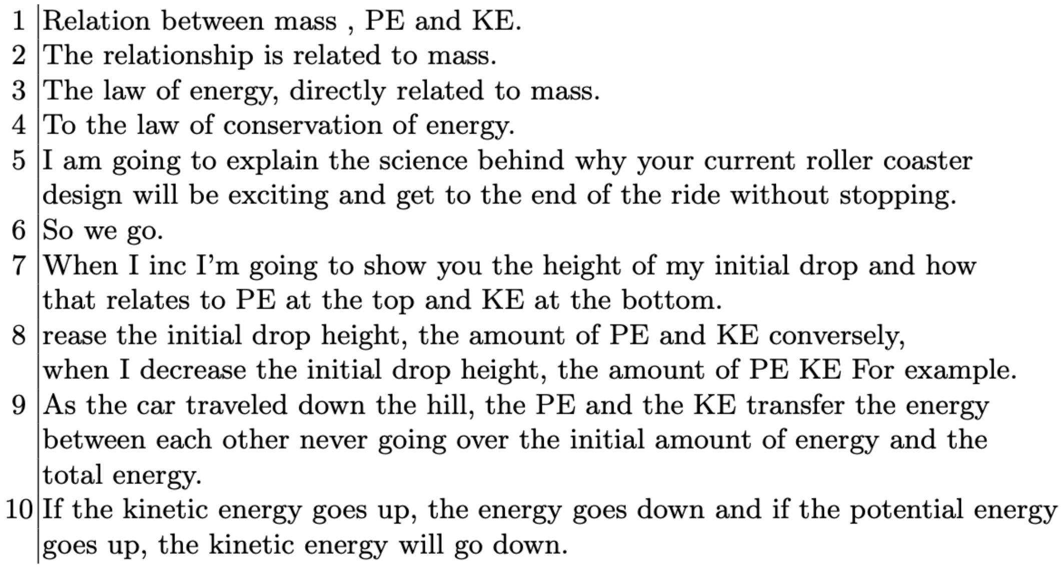
Table3: Accuracy on Main Ideas (as percentages)

The distinctiveness of the main ideas plays a crucial role in determining its accuracy in PyrEval. When a main idea is distinctive and definitive, such as the law of conservation, PyrEval tends to exhibit higher accuracy in its assessments. This is because such concepts have a well-established, narrow range of acceptable explanations, making it easier for PyrEval to accurately evaluate a student's comprehension. On the other hand, concepts that are more open to interpretation or can be expressed in a variety of ways, like the relationship between potential and kinetic energy, often result in lower accuracy scores. Therefore, the distinctiveness of the main ideas can significantly impact the accuracy of PyrEval.

Chapter 6
Analysis of Student Writing and Idea Distinctiveness

In this chapter, we examine a second important factor, one that we later relate to the opportunity for students to reflect on their expressive skills: student writing clarity.

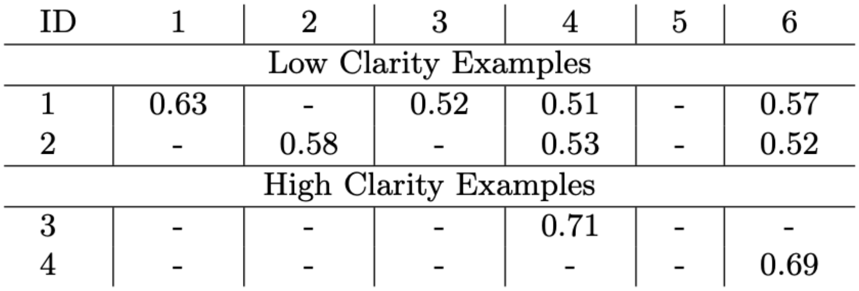
On the one hand, clarity of student writing is fairly obvious to adult readers, as illustrated in the sample essay shown in Figure2. Given its punctuation, the essay is processed as having ten sentences. Sentences 1-4 and 6 are not full sentences, and 5 merely states the purpose of the essay. The final sentences are much more complete, and fairly clear. (It seems possible that the student mixed up material across sentences 7 through 9, perhaps through faulty cut-and-paste steps.) On the other hand, clarity of writing results from many factors, some of which would be difficult to measure. Instead of attempting to directly measure writing clarity, we assess differences across student essays using cosine similarity of student phrases to main ideas as a probe.



1 Relation between mass , PE and KE.
2 The relationship is related to mass.
3 The law of energy, directly related to mass.
4 To the law of conservation of energy.
5 I am going to explain the science behind why your current roller coaster design will be exciting and get to the end of the ride without stopping.
6 So we go.
7 When I inc I'm going to show you the height of my initial drop and how that relates to PE at the top and KE at the bottom.
8 rease the initial drop height, the amount of PE and KE conversely, when I decrease the initial drop height, the amount of PE KE For example.
9 As the car traveled down the hill, the PE and the KE transfer the energy between each other never going over the initial amount of energy and the total energy.
10 If the kinetic energy goes up, the energy goes down and if the potential energy goes up, the kinetic energy will go down.

Figure2: A student essay with very mixed writing quality

For further data analysis in this chapter, we randomly chose 117 essays out of 159, using the remaining 42 essays to verify the consistency of our findings. We categorized the essays into three groups based on the number of PyrEval errors: those with one or fewer errors (58 essays; High Accuracy), those with two errors (45 essays; Medium Accuracy), and those with more than two errors (14 essays; Low Accuracy). After that, the cosine similarity of vectors of student clauses and vectors of main idea clauses in the pyramid content units are calculated.



ID	1	2	3	4	5	6
Low Clarity Examples						
1	0.63	-	0.52	0.51	-	0.57
2	-	0.58	-	0.53	-	0.52
High Clarity Examples						
3	-	-	-	0.71	-	-
4	-	-	-	-	-	0.69

Figure3: Clauses with low versus high clarity, and main ideas they are similar to

The examples in Figure3 illustrate the same point about the impact on accuracy if a clause has above threshold similarity to more than one main idea. The top of the figure shows two phrases that are poorly written, and that have cosine similarities above t for multiple ideas. The lower half of the figure shows two well-articulated statements, where each is above threshold similarity to exactly one main idea, and also where the cosine similarity is much higher than 0.50.

A notable observation emerges when comparing essays with perfect accuracy and shorter length (100% accuracy, average length ~350 words) against those with lower accuracy and longer length (50% accuracy, length > 500 words) as depicted in Figure4, Figure5 and Figure6. The accurate essay (darker bars) has a lower count of clauses overall, but more importantly, very few that have a cosine similarity of 0.70 and above. In contrast, the inaccurate essay has about ten times as many at that cosine similarity and above, which increases the chances that the node selected by the MIS algorithm as a clause matching a main idea would not be one that a human would select.

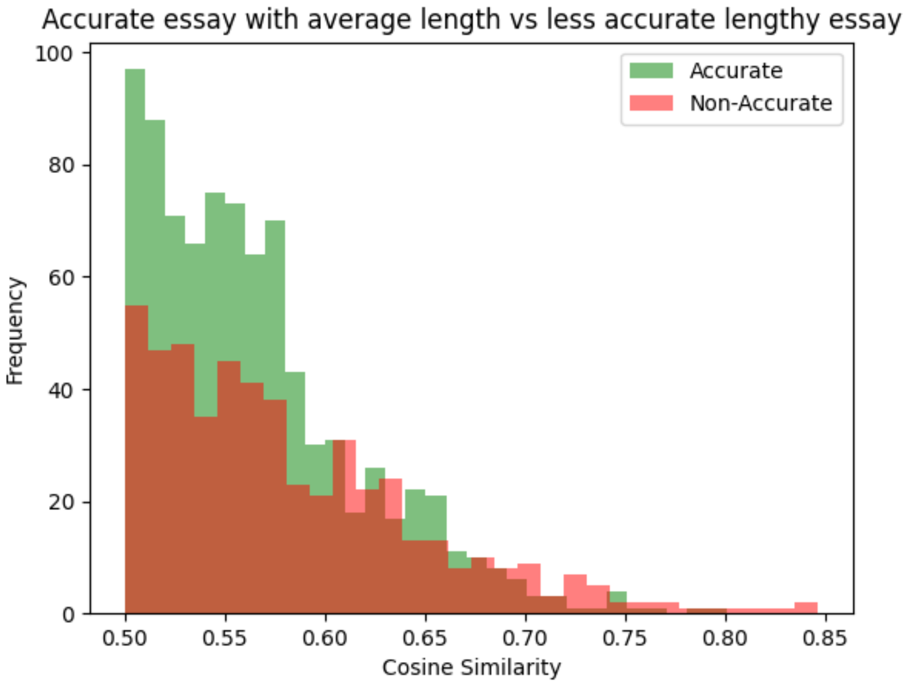


Figure4: Cosine Similarity distribution of clauses in the full assessment hypergraph for an accurate short essay, and long inaccurate essay, example1

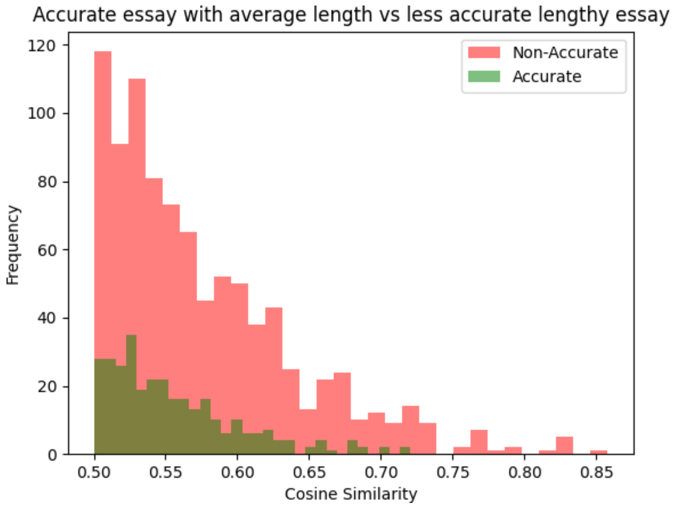


Figure5: Cosine similarity distributions of clauses in the full assessment hypergraph for an accurate short essay, and a long inaccurate essay, example 2

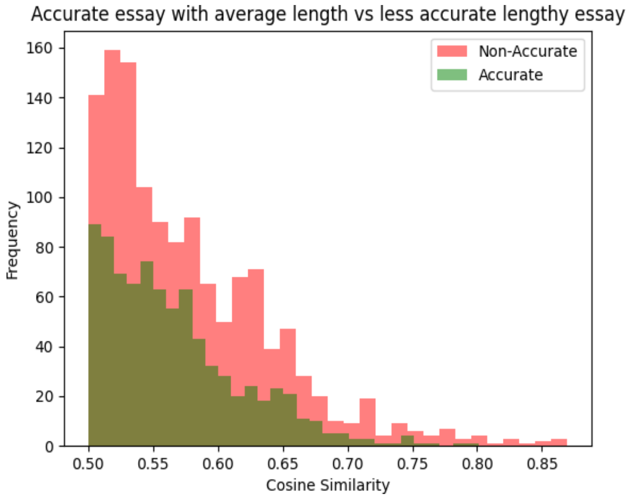


Figure6: Cosine similarity distributions of clauses in the full assessment hypergraph for an accurate short essay, and a long inaccurate essay, example 3

When analyzing essays based on the accuracy of content units, as shown in Figure7, Figure8 and Figure9, several trends become apparent. The cosine similarities in accurate CUs display a stair-step-like progression, while in inaccurate CUs, high cosine values drop more dramatically in frequency. By the shape of its plot, it is evident that the content unit related to the law of conservation presents a unique case, setting it apart from other content units and marking it an outlier.

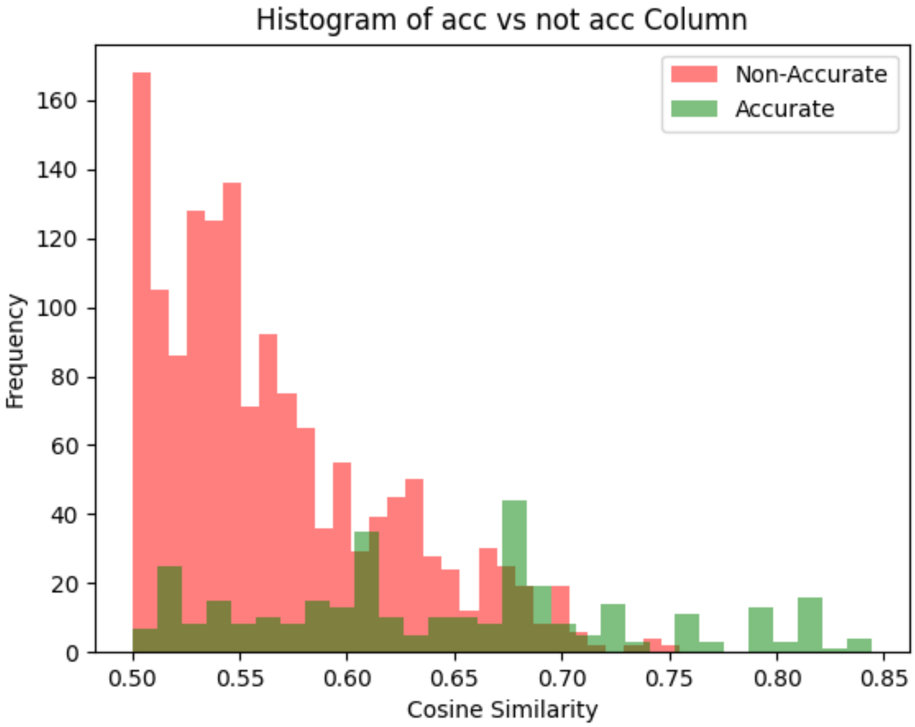


Figure7: Main Idea 1 (not accurate) vs Main Idea 4 (accurate)

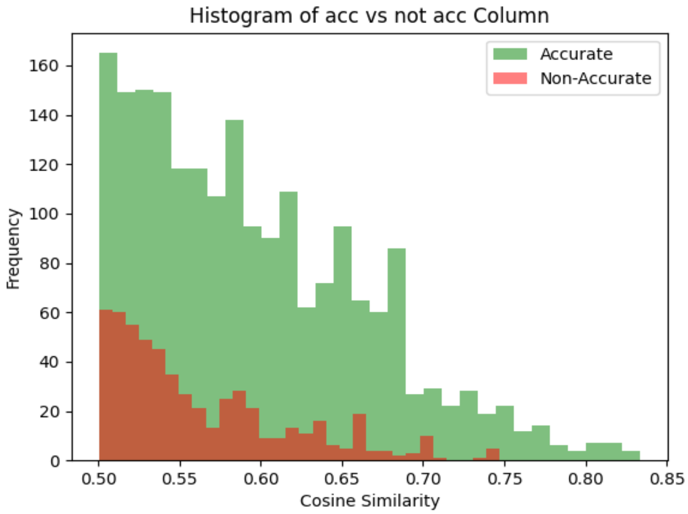


Figure8: Main Idea 3 (not accurate) vs Main Idea 5 (accurate)

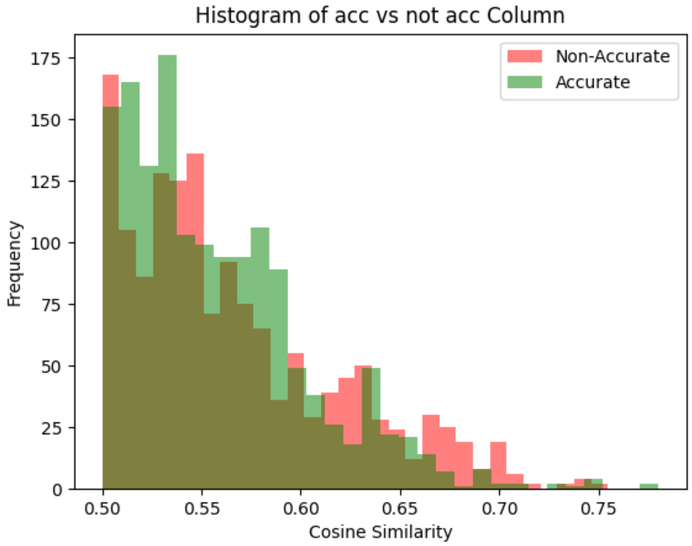


Figure9: Main Idea 1 (not accurate) vs Main Idea 6 (accurate)

When considering PyrEval accuracy, we examined sets of essays that had high overall accuracy versus low accuracy by comparing the cosine similarities of phrases from the accurate essays across all main ideas to a given main idea with the cosine similarities of phrases from the inaccurate essay to the same main idea. The distributions were clearly different for the high versus low accuracy sets of essays, and also were different depending on the main idea. For a given main idea, accurate essays generally show phrases that have very high cosine similarity values above 0.7, meaning a longer right tail, and fewer phrases that have moderate cosine similarity values, meaning between 0.5 and 0.7. This pattern is illustrated in Figure12 and Figure13 below, for main ideas 1 and 3, for the accurate essays. In contrast, the same figures show that the inaccurate essays display flatter distribution without a long right tail.

Given the way WMIN arrives at a final assessment, these figures suggest how the different distributions of cosine similarities lead to high versus low accuracy. For each main idea, WMIN tries to select an essay from a student phrase that has the highest cosine similarity to the corresponding content unit, while also assigning the highest overall sum of CU weights. If there are only a very few phrases that have a very high cosine similarity to a given main idea, WMIN is more likely to select the correct one. This is the pattern we observe for the accurate essays, for main ideas 1 and 3 (see Figure12 and Figure13). However, if there are many phrases that have a moderate cosine similarity to a given main idea, WMIN is less likely to select a correctly matching phrase. This is the pattern we observe in the same figures, for the inaccurate essays.

Furthermore, in another type of plot we noticed a difference between high versus low accuracy essays that illustrates a different kind of WMIN error. Accuracy can be broken down into positive accuracy (sensitivity) and negative accuracy (specificity). The next set of plots are cosine similarity distributions for essays that had high versus low negative accuracy plots. It appears that for some main ideas (etc. MI1 and MI3), as you can see in Figure10 and Figure11, there are higher cosine similarity values for the inaccurate ones than the accurate ones, making it very likely PyrEval will select these high-cosine similarity segments, even when they do not correspond to the main idea in question.

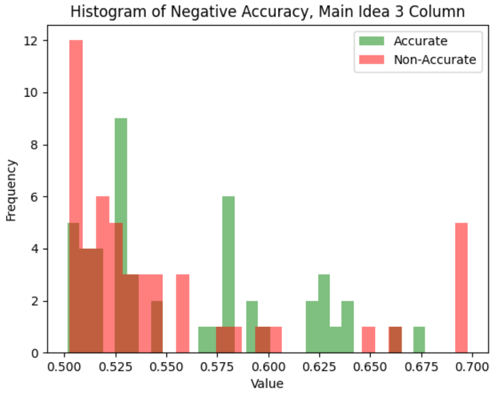


Figure10: Negatively accurate vs negatively not accurate on main idea 3

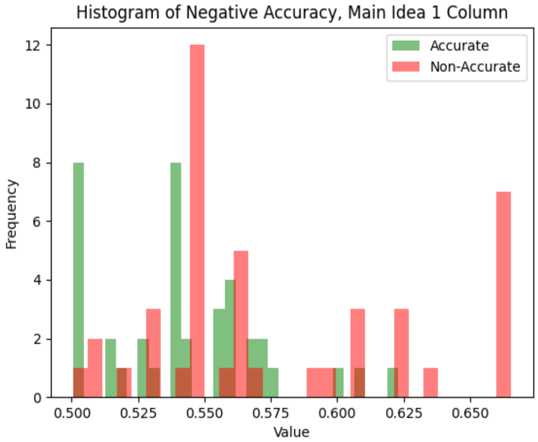


Figure11: Negatively accurate vs negatively not accurate on main idea 1

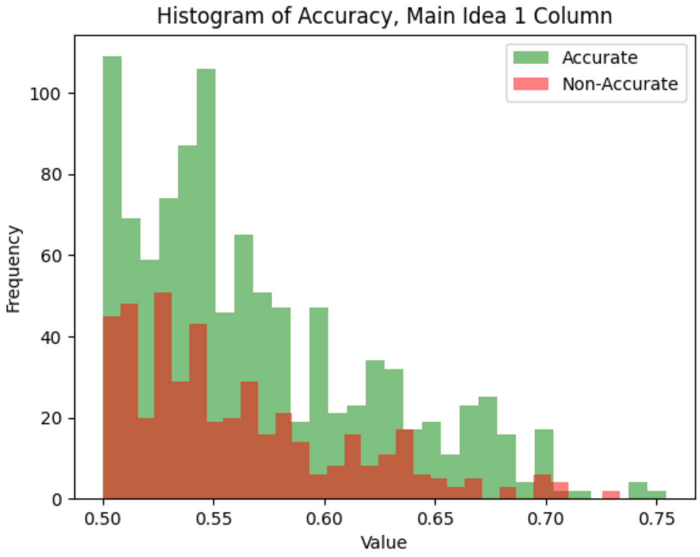


Figure12: Overall accurate vs overall not accurate on main idea 1

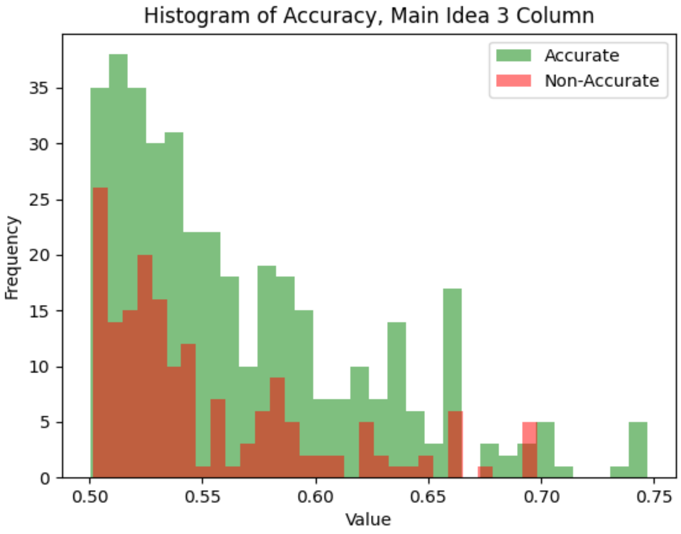


Figure13: Overall accurate vs overall not accurate on main idea 3

In this part, we categorized the essays based on overall essay accuracy (high accuracy>0.8, 0.6<mid accuracy ≤0.7, low accuracy≤0.6), depicted in Figure14, Figure15 and Figure16, to differentiate them in cosine similarity distribution plots based on each main idea. Interestingly, mid and high accuracy essays demonstrate similar patterns for lower cosine similarity values (below 0.7). In addition, almost no cosine similarity values QA from the low accuracy essays is observed beyond 0.8 on the plots.

Essays with high accuracy feature most of the very high cosine similarity values (above 0.7) with a more curve-like distribution. In contrast, low accuracy essays exhibit a flatter distribution of cosine values, indicating a lack of clarity in the students' writing.

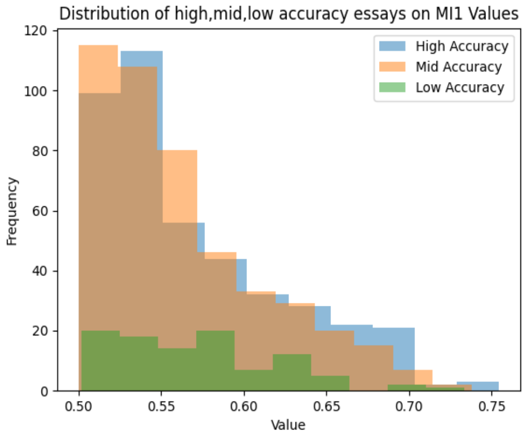


Figure14: Distribution of cosine similarities in high, mid and low accuracy essays on main idea 1

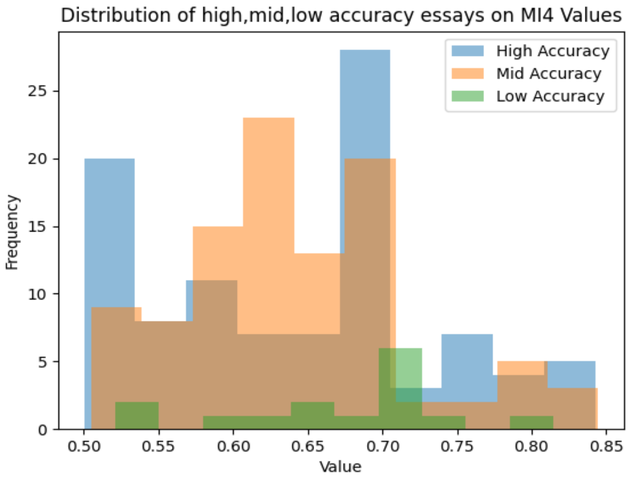


Figure15: Distribution of cosine similarities in high, mid and low accuracy essays on main idea 4

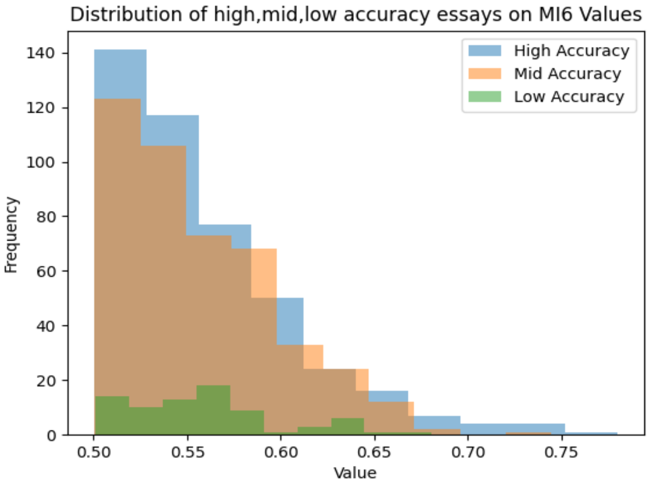


Figure16: Distribution of cosine similarities in high, mid and low accuracy essays on main idea 6

In general, plots of the cosine similarity distributions of student clauses to main ideas have a pattern where the high accuracy essays have distributions that extend to relatively higher cosine values of 0.70 and above but with low counts. For the lower range of cosine similarities values between 0.50 and 0.70, there are typically many clauses. In contrast, the distribution of cosine similarities to main idea 4, the law of conservation of energy, is quite distinctive.

Chapter 7
Comparison of PyrEval with Alternative Methods

In this chapter, we compare PyrEval to an alternative method called VerAs, which employs an end-to-end neural architecture. This analysis highlights the differences in data requirements and overall accuracy, providing insights into the strengths and limitations of each approach in automated content assessment.

VerAs

VerAs is an end-to-end neural architecture that has separate verifier and assessment modules, inspired by approaches to Open Domain Question Answering (OpenQA). VerAs first verifies whether a report contains any content relevant to a given rubric dimension, and if so, assesses the relevant sentences. VerAs was developed to apply multi-dimension analytic rubrics to physics lab reports from an introductory inquiry-based undergraduate course. As shown in Figure17, VerAs has two modules, a Verifier module to process the rubric dimension, analogous to a query in OpenQA, and a Grader module to determine the assessment result. The Verifier decides whether the report has a non-zero score, and if so, selects the top ***k*** sentences to pass to the Grader, where ordinal log loss is used to determine what grade on a 6-point scale to assign. My contribution was to test the Verifier module on the middle school essays.

We conduct experiments on data for two middle school essay assignments, along with analytic rubrics for formative feedback, and where each rubric has a different number of dimensions (six for essay 1; eight for essay 2).

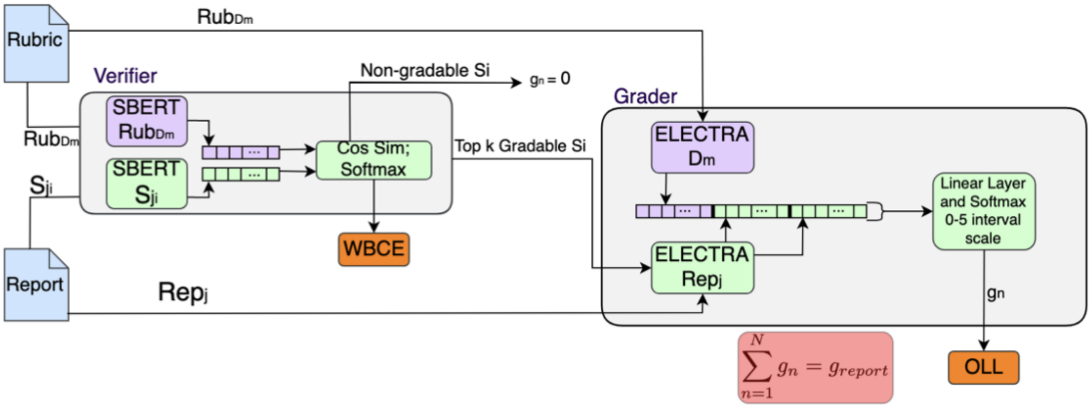
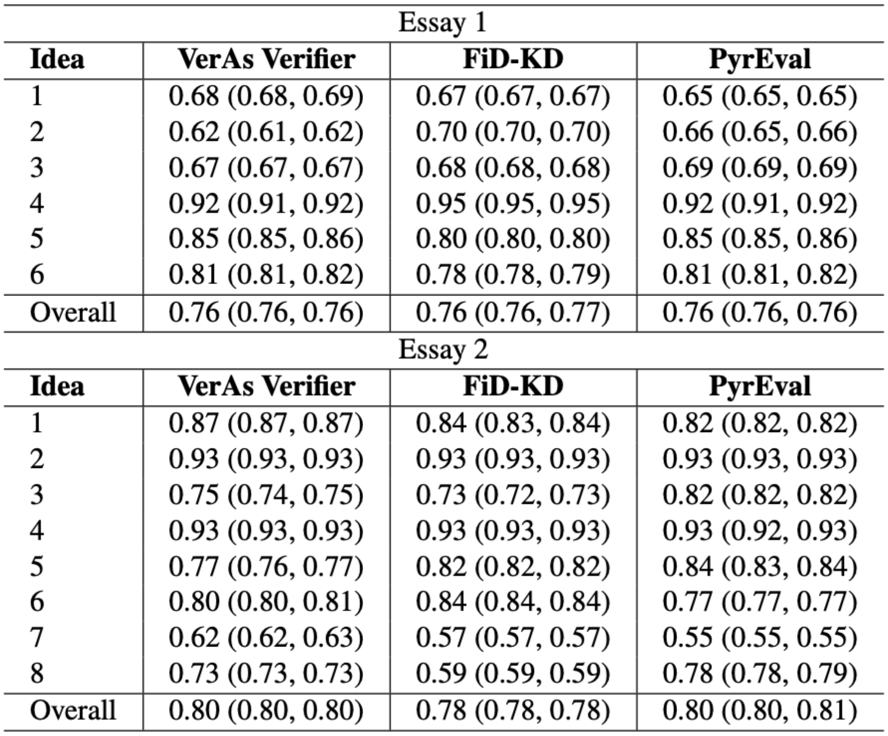


Figure17: VerAs: Using a dual encoder, the verifier assesses each report sentence (Si) and rubric dimension (Dm) to forward the top k sentences to the grader, trained with weighted binary cross-entropy loss on whether the report receives a non-zero score.

Each essay rubric dimension is an explanatory statement of one of the main ideas in the curriculum. These can be more general, such as how potential and kinetic energy in a roller coaster are related to one another, or more specific, such as an explanation of the law of conservation of energy. Instead of assessing each dimension on a scale, the essay feedback indicates only whether the student included a clear statement of one of the main ideas. As a result, the VerAs grader module plays no role.

Only 159 of the essays have reliable manual labels indicating the presence of main ideas (Cohen’s kappa = 0.77) (essay 1 test is entirely manual labels). The remaining labels are from PyrEval whose accuracies on the two essays are 0.76 and 0.80, respectively, as you can see in Table4. Thus, VerAs is trained on noisy data. For essay 2, there are reliable manual labels on 56 essays, corresponding to the essay 2 test set. FiD-KD is the OpenQA system that most directly inspires VerAs and is used as a baseline.



Essay 1

Idea	**VerAs Verifier**	**FiD-KD**	**PyrEval**
1	0.68 (0.68, 0.69)	0.67 (0.67, 0.67)	0.65 (0.65, 0.65)
2	0.62 (0.61, 0.62)	0.70 (0.70, 0.70)	0.66 (0.65, 0.66)
3	0.67 (0.67, 0.67)	0.68 (0.68, 0.68)	0.69 (0.69, 0.69)
4	0.92 (0.91, 0.92)	0.95 (0.95, 0.95)	0.92 (0.91, 0.92)
5	0.85 (0.85, 0.86)	0.80 (0.80, 0.80)	0.85 (0.85, 0.86)
6	0.81 (0.81, 0.82)	0.78 (0.78, 0.79)	0.81 (0.81, 0.82)
Overall	0.76 (0.76, 0.76)	0.76 (0.76, 0.77)	0.76 (0.76, 0.76)

Essay 2

Idea	**VerAs Verifier**	**FiD-KD**	**PyrEval**
1	0.87 (0.87, 0.87)	0.84 (0.83, 0.84)	0.82 (0.82, 0.82)
2	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)
3	0.75 (0.74, 0.75)	0.73 (0.72, 0.73)	0.82 (0.82, 0.82)
4	0.93 (0.93, 0.93)	0.93 (0.93, 0.93)	0.93 (0.92, 0.93)
5	0.77 (0.76, 0.77)	0.82 (0.82, 0.82)	0.84 (0.83, 0.84)
6	0.80 (0.80, 0.81)	0.84 (0.84, 0.84)	0.77 (0.77, 0.77)
7	0.62 (0.62, 0.63)	0.57 (0.57, 0.57)	0.55 (0.55, 0.55)
8	0.73 (0.73, 0.73)	0.59 (0.59, 0.59)	0.78 (0.78, 0.79)
Overall	0.80 (0.80, 0.80)	0.78 (0.78, 0.78)	0.80 (0.80, 0.81)

Table4: Accuracy on Main Ideas (as percentages)

VerAs, similar to PyrEval, performs well on an analytic rubric for middle school physics essays. While PyrEval can create content models from as few as 4 or 5 reference passages, and requires no training data, VerAs requires a good amount of reliable training data which may be difficult to collect.

Chapter 8
Conclusion

In this study, I examined various aspects of PyrEval, an off the shelf automated assessment tool, and the factors that affect the accuracy of it. In addition to that, this thesis focused on the comparative performance of PyrEval and VerAs. Our findings highlight several key insights into how these tools function and the variables that influence their effectiveness.

Factors Influencing Accuracy:

1- Curriculum Ideas and Distinctiveness: The accuracy of our automated assessment tool is influenced by the distinctiveness of curriculum ideas. Curriculum ideas with clear and unique characteristics are easier for the tool to identify and assess accurately. Conversely, less distinctive ideas can lead to ambiguities, reducing the tool's accuracy.

2- Clarity of Student Articulation: The way students articulate their ideas also impacts the accuracy of the assessment. Clear and well-structured statements are more likely to be accurately assessed by the tool. In contrast, vague or poorly articulated statements pose a challenge, potentially leading to inaccurate assessments.

In another part of this thesis, I examined that PyrEval has demonstrated an advantage in its robustness to non-standard writing features. It does not penalize essays for unconventional writing styles, allowing it to assess content without being affected by linguistic variations. This flexibility makes PyrEval particularly useful in diverse educational settings where students may express their ideas in non-traditional ways.

In comparing PyrEval with VerAs, another assessment method, we concluded that VerAs requires extensive training data to achieve the same high accuracy as PyrEval. Unlike PyrEval, VerAs relies heavily on large datasets to train its models, which can be a limitation in contexts where such data is not readily available.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423

2. Gao, Y., Sun, C., Passonneau, R.J.: Automated pyramid summarization evaluation. In: Bansal, M., Villavicencio, A. (eds.) Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 404–418. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/K19-1038, https://aclanthology.org/K19-1038

3. Panadero et al., 2023 -- meta-analysis review: Panadero, E., Jonsson, A., Pinedo, L., Fernández-Castilla, B.: Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: a meta-analytic review. Educational Psychology Review 35, article 113 (2023). https://doi.org/10.1007/s10648-023-09823-4

4. Tansomboon, C., Gerard, L.F., Vitale, J.M., Linn, M.C.: Designing auto- mated guidance to promote productive revision of science explanations. International Journal of Artificial Intelligence in Education 17, 729–757 (2017). https://doi.org/10.1007/s40593-017-0145-0

5. Gerard, L., Linn, M.C., Madhok, J.: Examining the impacts of annotation and automated guidance on essay revision and science learning. In: Looi, C.K., Polman, J.L., Cress, U.,

Reimann, P. (eds.) Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) (2016)

6. Gerard, L., Linn, M.C.: Computer-based guidance to support students' revi- sion of their science explanations. Computers & Education 176, 104351 (2022). https://doi.org/10.1016/j.compedu.2021.104351

7. Heilman, M., Madnani, N.: ETS: Domain adaptation and stacking for short answer scoring. In: Manandhar, S., Yuret, D. (eds.) Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceed. of the 7th International Workshop on Semantic Evaluation (SemEval 2013). pp. 275–279. Assoc. for Computational Linguistics, Atlanta, GA (Jun 2013), https://aclanthology.org/ S13- 2046

8. Lee, H.S., Gweon, G.H., Lord, T., Paessel, N., Pallant, A., Pryputniewicz, S.: Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. Journal of Science Education & Technology pp. 168–192 (2021). https://doi.org/10.1007/s10956-020-09889-7

9. Zhu, M., Liu, O.L., Lee, H.S.: The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. Computers & Education 143, 103668 (2020). https://doi.org/10.1016/j.compedu.2019.103668

10. Lee, H.S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., Liu, O.L.: Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. Science Education 103(3), 590–622 (2019). https://doi.org/https://doi.org/10.1002/sce.21504

11. Pallant, A., Lee, H.S., Pryputniewicz, S.: How to support secondary school stu- dents' consideration of uncertainty in scientific argument writing: A case study of a high-adventure science curriculum module. Journal of Geoscience Education 68(1), 8–19 (02 2020)

12. Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsmura, L.C., Howe, E., R., Q.: eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In: Thirty-First AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-19) (2019).https://doi.org/10.1609/aaai.v33i01.33019619

13. Singh, P., Passonneau, R.J., Wasih, M., Cang, X., Kim, C., Puntambekar, S.: Automated Support to Scaffold Students' Written Explanations in Science. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education, vol. 13355, pp. 660–665. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-11644-5 64

14. Weiwei Guo and Mona Diab. 2012. Modeling Sentences in the Latent Space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 864–872, Jeju Island, Korea. Association for Computational Linguistics.

15. 1. Kučera H. Francis W. N. (1967). Computational analysis of present-day American English Providence, R.I.: Brown University Press.

16. Karizaki, M.S., Gnesdilow, D., Puntambekar, S., Passonneau, R.J. (2024). How Well Can You Articulate that Idea? Insights from Automated Formative Assessment. In: Olney,

A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science(), vol 14830. Springer, Cham. https://doi.org/10.1007/978-3-031-64299-9_16

17. Atil, B., Sheikhi Karizaki, M., J. Passonneau, R. (2024). VerAs: Verify Then Assess STEM Lab Reports. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science(), vol 14829. Springer, Cham. https://doi.org/10.1007/978-3-031-64302-6_10

18. Kim, C., Puntambekar, S., Lee, E., Gnesdilow, D., Karizaki, M. S., & Passonneau, R. (2024). NLP-Enabled Automated Feedback about Science Writing. In Lindgren, R., Asino, T. I., Kyza, E. A., Looi, C. K., Keifert, D. T., & Suárez, E. (Eds.), Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024 (pp. 2431-2432). International Society of the Learning Sciences.