Cambium Assessment

Write On with Cambi: The development of an argumentative writing feedback tool

Susan Lottridge, Amy Burkhardt, Christopher Ormerod, Sherri Woolf, Mackenzie Young, Milan Patel, Harry Wang, Julius Frost, Kevin McBeth, Julie Benson, Michael Flynn, Kevin Dwyer, Scott Fitz, Radd Berkheiser, Henry Floyd, Dave Davis, Ben Godek, Quinell Wilson

Cambium Assessment, Inc.

Author Note

Correspondence concerning this article should be addressed to Sue Lottridge, Cambium Assessment, Inc., Email address: susan.lottridge@cambiumassessment.com

Abstract

Every year, millions of middle-school students write argumentative essays that are evaluated against a scoring rubric. However, the scores they receive don't necessarily offer clear guidance on how to improve their essay or what they've done well. With advancements in natural language processing technology, we now have the capability to provide more detailed feedback. At this juncture, we've developed an artificial intelligence-supported editing tool to assist students in revising their essays. In this paper, we introduce this tool. We then delve into its underlying components, covering how the feedback aligns with the rubric and standards, the techniques employed in modeling, and the process of crafting feedback. The paper concludes by summarizing the results obtained from interviews with teachers who have used the tool.

Write On with Cambi!: The Development of an Argumentative Writing Feedback Tool

In the United States, every year millions of middle school students write argumentative essays which are then evaluated according to a specific set of criteria. These criteria are standardized across different assessment rubrics, including the rubrics for *Smarter Balanced* (2022) and *Integrating College and Career Readiness* (ICCR) items, which are used by many clients of Cambium Assessment[1]. This standardization ensures consistency in assessing writing proficiency across various score levels. These rubrics are used for formative, interim, and summative assessments, with scores assigned by educators, a hand-scoring vendor, or through a machine learning model. While these scores serve the key purpose of providing high level evaluations of writing, they often fail to provide insights that can guide students in becoming better writers. This limitation arises from the inherent nature of rubrics, wherein essays within a score level may exhibit a range of writing characteristics. Indeed, writing is a complex, multi-faceted task for which a single score or set of trait-level scores simply cannot reflect the breadth of writing characteristics.

In the absence of an educator serving as a mediator between the rubric and the student, it becomes difficult for a student to decipher how their rubric score can inform the next steps in their writing process. It is also time-consuming for an educator to review each student essay against a rubric score and identify areas that need to be improved or that are working well. However, the recent technological advancements responsible for improving automated scoring accuracy also enables more precise detection of finer-grained elements of writing (e.g., Lottridge et al., 2023; Kaggle, 2022). This capability is a prerequisite for delivering standards-aligned and automated feedback to students.

We have leveraged this new technology to develop an online editing tool, called "Write On with Cambi!", targeted towards helping students to review their essays. In this paper, we describe how "Write On with Cambi!" that detects both argumentative elements in student writing as well as conventions-related errors and provides targeted feedback. Our goal with "Write On with Cambi!" is to provide high-quality standards-based feedback that helps students as they review their essay and aligns with how educators teach writing in their classrooms. Our hope is that such a tool can support educators and students in the revising enterprise, so that students produce high-quality essays and – with practice – become better writers.

This paper details the behind-the-scenes development of this tool, aiming to provide evidence that the finer-grained inferences drawn by the tool accurately reflect students' writing abilities and are aligned with what educators value and teach. The structure of this paper is as follows. The editing tool is first introduced, and then the subsequent sections detail the development of the tool (focused primarily on labeling the data and modeling techniques). The final section summarizes results from teachers' reactions to interacting with "Write On with Cambi!".

[1] Note that there are some nuanced differences in the language used to describe various criteria across these two rubrics, but these differences do not result in qualitatively different distinctions across the criteria within each level of the rubric.

Write On with Cambi!

This tool, developed by Cambium Assessment, Inc., is referred to as *Write On with Cambi!* It is embedded into the organization's test delivery system (Figure 1).

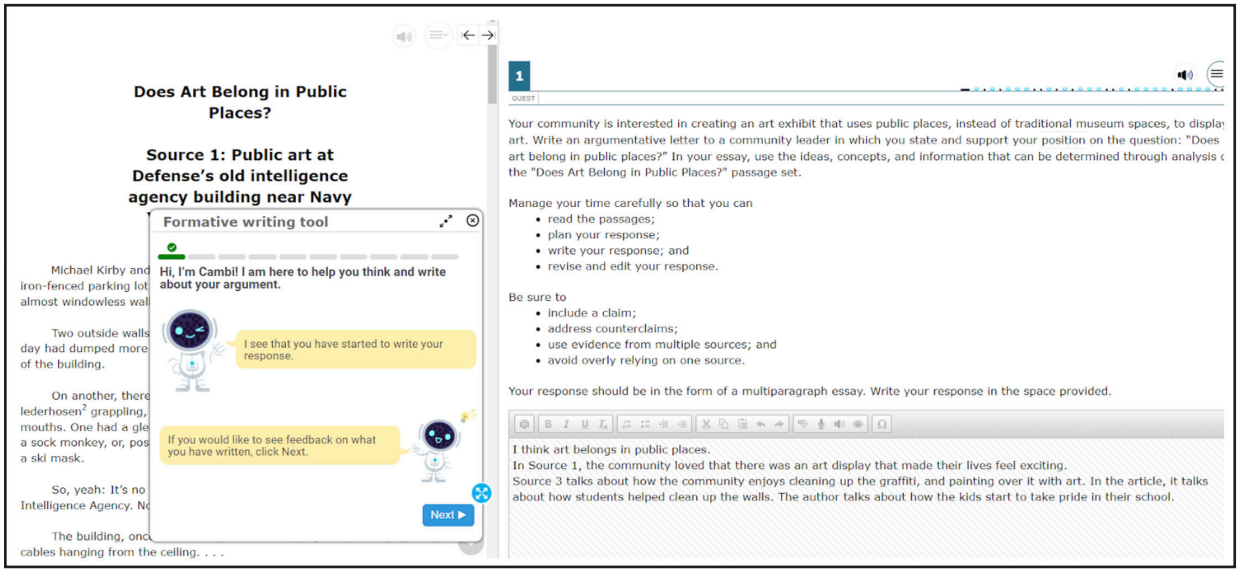

Figure 1. The Introduction Slide of the Writing Tool

After a student has drafted an argumentative essay, they can request assistance from Cambi, the robot, to guide them through the editing process. As of now, the tool is targeted for students in grades 6 through 8 for writing argumentative essays. This group was chosen because research showed that the annotation of argumentative elements could be modeled (Kaggle) and because this grade band of students was considered to benefit most from a writing feedback tool.

The feedback provided by the tool is organized into two distinct sets. The first set of feedback pertains to the argumentative elements of the essay; the second set addresses writing conventions. This sequencing was purposeful: namely, we wanted the student to focus first on the substantive elements of argumentative writing, and then focus on the editing aspects around conventions. The sequencing of the feedback associated with the key argumentative elements is arranged into the three phases. First, Cambi, examines the organization of the essay, and checks for the presence or absence of text indicating an Introduction, Conclusion, and Main Claim (Figure 2).

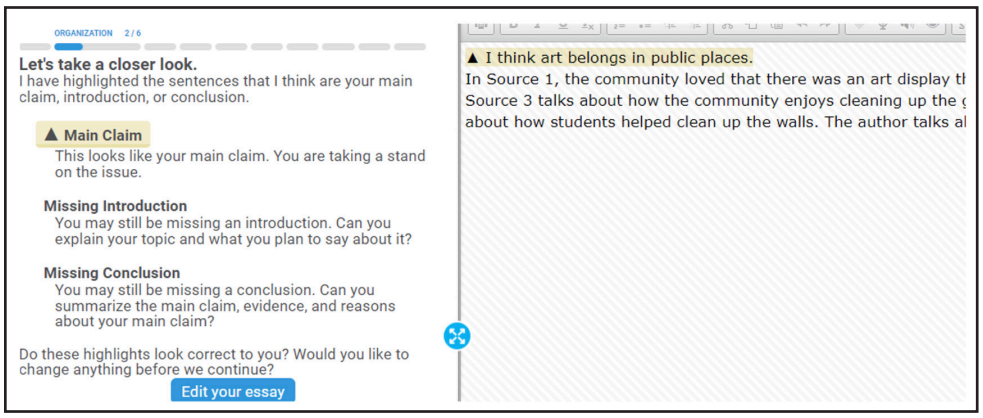

Figure 2. A Slide Related to the Organization of the Essay within the Writing Tool

Then, Cambi guides the student through reviewing their Evidence and Reasons to support their claim (Figure 3). In this section, Cambi not only examines the presence or absence of *Evidence* and *Reasons*, but also examines if they are integrated.

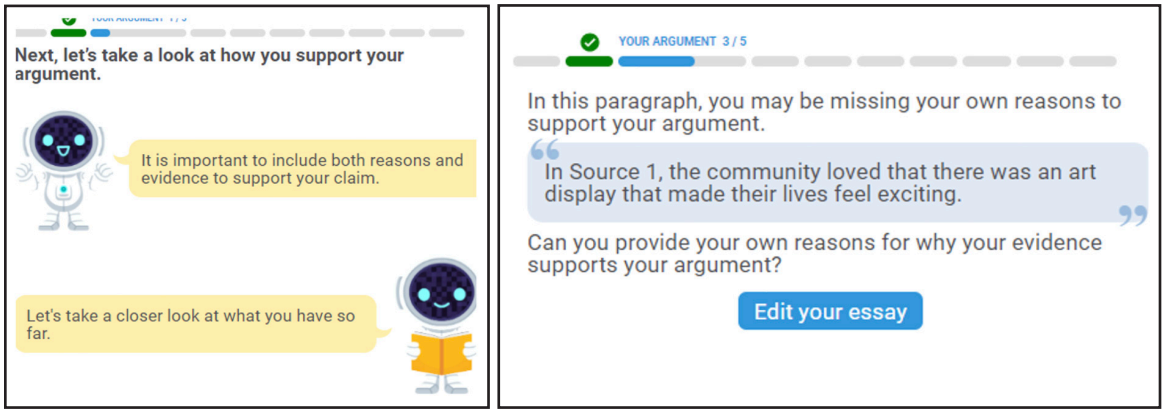

Figure 3. Two Slides within the Tool Related to the Evidence and Reasons

Finally, Cambi reminds the student to not only include an Opposing Position, but to also include a *Rebuttal* (Figure 4).

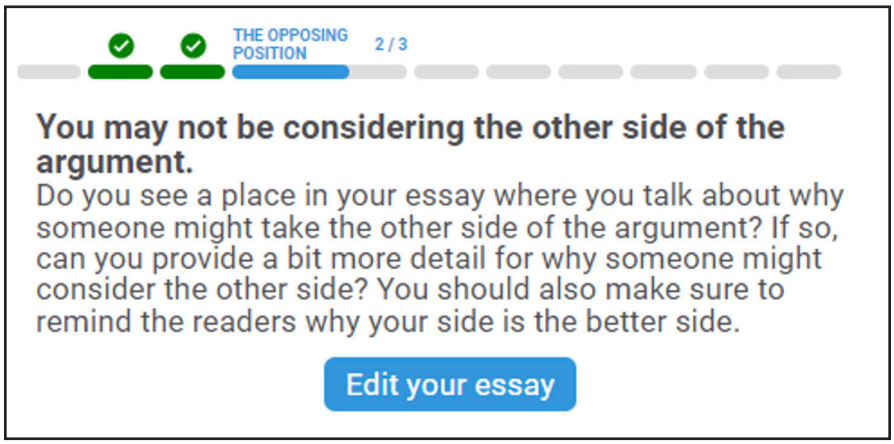

Figure 4. One Slide Related to the Opposing Position and Rebuttal

Students are also provided with an overview of the annotations, both at the onset of the editing process and upon completion, enabling them to reflect on the revisions that they have implemented (Figure 5 and Figure 6).

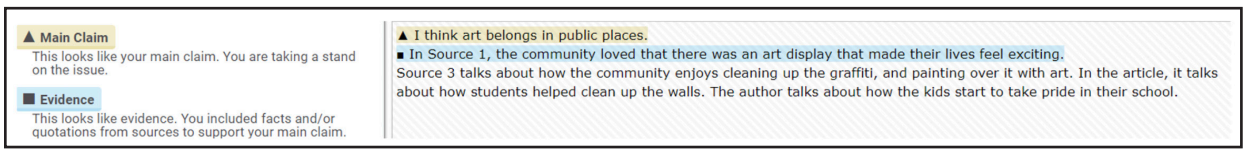

Figure 5. An Example of the Overview of Annotation Elements at the Onset of Editing

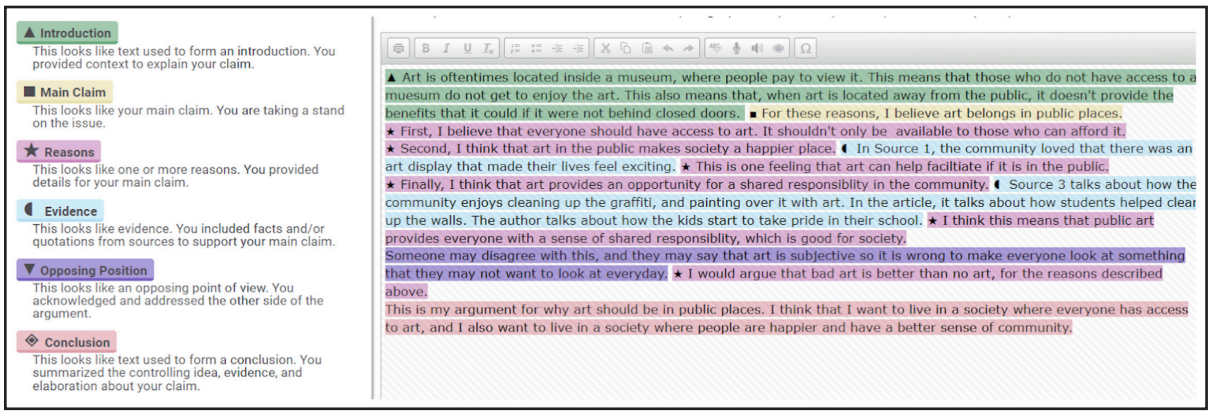

Figure 6. An Example of the Overview of Annotation Elements upon Completion of Feedback

The conventions section then follows, and attends to errors relating to Sentence Formation, Capitalization, Punctuation, Spelling, and Grammar. These groupings reflect those emphasized the Smarter Balanced and ICCR rubrics. In this section, corrections are never automated for the student; rather, students are guided to correct the error themselves (Figure 7).

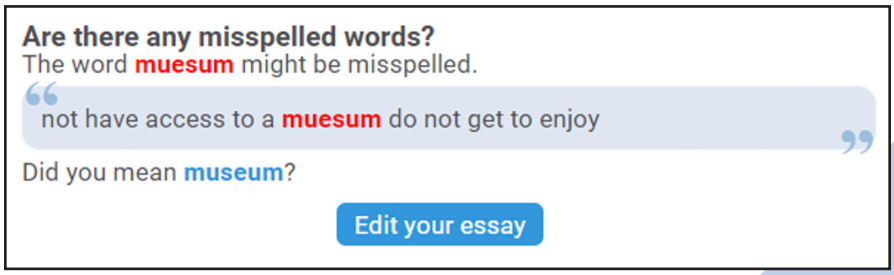

Figure 7. Example of the Tool Guiding the Student to Make a Correction

The following section outlines the procedure for aligning the argumentative elements and the conventions-related components with the two primary rubrics and state standards that act as a framework for this tool. Furthermore, the section details the labeling process for both the argumentative elements and the *Sentence Formation* subcategories under conventions.

Aligning Finer-Grained Elements to the Rubric and Developing Annotation Guidelines

In this section, we revisit the criteria in the aforementioned rubrics that serve as the framework for establishing guidelines to (a) determine the specific finer-grained elements for which the tool will offer feedback, and (b) label data essential for training various machine learning models. As exhibited in Appendix A, the rubrics consist of three dimensions: (1) Purpose, Focus, and Organization, (2) Evidence and Elaboration, and (3) Conventions of Standard English.

Argumentation Elements

As outlined in the two dimensions of the rubric – *Purpose, Focus, and Organization* and *Evidence and Elaboration* – a well-formed argumentative essay must incorporate the following elements: *Introduction*, *Conclusion*, *Main Claim*, *Opposing Position*, *Transitions*, *Evidence* and *Reasons*. These seven elements formed the basis for the annotation guidelines crafted to assist hand-scorers in labeling data used to train and validate a machine learning engine. Figure 8 illustrates a visual depiction of a fully annotated essay, with each sentence labeled with a specific argumentation element.

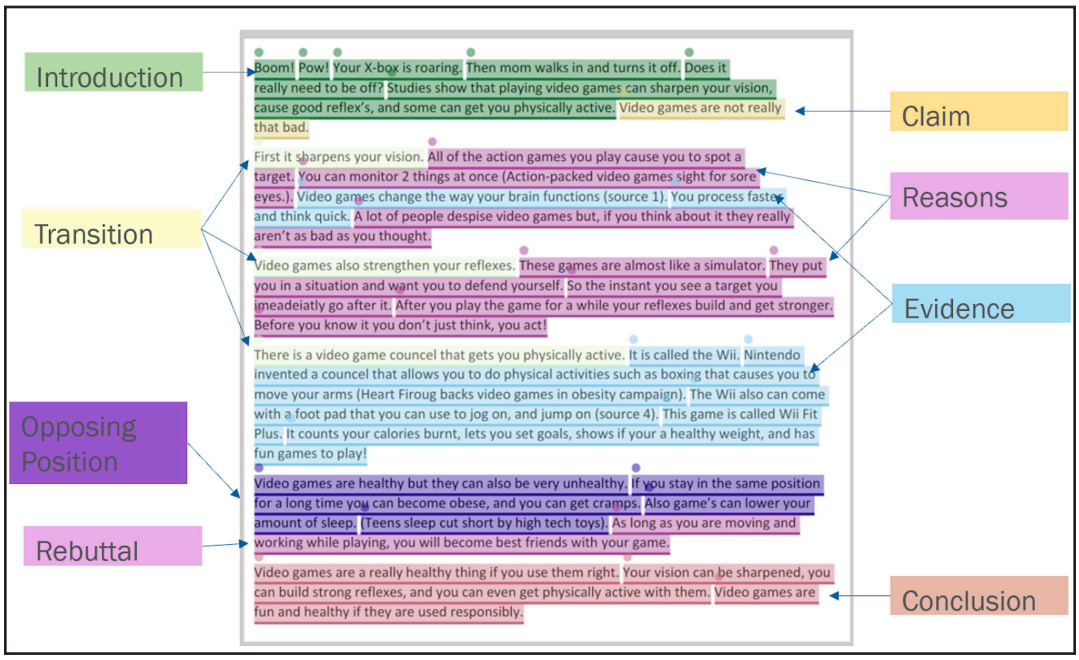

Figure 8. An Example of an Essay Annotated for its Argumentative Elements

To ensure consistent labeling by hand scorers, we developed guidelines (Appendix B) outlining the criteria for annotations, including the specific text or content that should be included and excluded for each argumentative element. The rubrics provided the foundational framework for these guidelines, which were subsequently honed using supplementary materials. It is important to highlight that the guidelines drew from the argumentative writing rubric applicable to 6th to 8th grade for the ICCR rubric and 6th to 11th grade for the Smarter Balanced rubric.

A rule was added to these guidelines, stipulating that only one tag should be assigned to each sentence. While this rule alleviates the cognitive load on raters by simplifying the process of applying multiple tags to a single sentence, it does present an issue when a sentence contains more than one argumentative element. Therefore, to ensure rater consistency, we implemented the following logic: If a sentence consists of two or more argumentative elements, assign the annotation tag that holds the highest priority within the following hierarchy: *Main Claim*, *Opposing Position*, *Evidence*, *Reasons*, *Transition*. The order of the elements within the hierarchy was motivated by ensuring that students are receiving credit for the most critical elements of an argumentative essay. This order was derived from the emphasis of each within the qualitative descriptors of the rubric.

Labeling the Argumentation Elements

The essays were labeled, or annotated, in the INCEpTION annotation software (Klie et al., 2018), which was hosted in a secure Amazon Web Service environment managed by CAI. The study utilized essays that were selected from responses written to nine prompts (three each in grades 6, 7, and 8) that were part of a summative assessment program in a Southern state of the United States. Responses were sampled from three academic years (2018-2019, 2020-2021, 2021-2022) to obtain data for each of the nine prompts.[2] Stratified random sampling was used

[2] 2020-2021 data were not used because the COVID-19 pandemic resulted in no summative testing that academic year. Note that this study did not examine the change in writing quality before and after the pandemic, as this was not a core concern of the study.

on the sum of the three rubric dimensions to ensure representation of all summed score points because the higher score points were rare in the student population. The online test administration for this state is supported by Cambium Assessment, Inc., and the total number of essays was 17,451 (approximately 2,000 for each of nine prompts). We chose this sample size and number of prompts to obtain a broad range of essays and prompts to reflect writing typical of students in grades 6 through 8.

Table 1 presents the score point distributions for the two dimensions of the rubric of interest for these annotations: The dimension of *Purpose, Focus, and Organization* (hereafter, Organization) and the dimension of *Evidence and Elaboration*. The distribution of scores for both dimensions are similar, in that very few responses received four points, and that most essays received two points across all nine prompts. An essay assigned scores of 2 points in both dimensions might indicate that, although it presents a main claim, its organizational structure may lack consistency. Additionally, while it may include both reasons and evidence, these elements may be weakly integrated.

Table 1. Score Point Distribution for Each Prompt (n = 17,451 Essays)

		Rubric Dimension: Organization				Rubric Dimension: Evidence and Elaboration			
Prompt ID	**Grade**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
A	6	38%	51%	10%	1%	65%	31%	4%	0%
B	6	46%	45%	8%	1%	62%	32%	5%	0%
C	6	44%	46%	10%	0%	58%	37%	5%	0%
D	7	16%	80%	4%	0%	35%	64%	2%	0%
E	7	35%	60%	5%	0%	56%	41%	3%	0%
F	7	36%	59%	4%	0%	54%	43%	3%	0%
G	8	26%	50%	22%	2%	28%	54%	17%	2%
H	8	23%	59%	17%	2%	28%	59%	12%	1%
I	8	30%	53%	16%	1%	37%	51%	12%	0%

A training session with thirteen experienced hand-scoring professionals, referred to as "annotators," occurred over a two-day period. These professionals were senior staff, responsible for training hand-scorers for large-scale summative scoring. Following this training, annotators were randomly assigned essays to annotate. This annotation took approximately two weeks, with some additional annotations occurring over two additional weeks to address differential rates across annotators. Fifteen percent of responses were randomly assigned to obtain a second annotation used to compute rater agreement. To evaluate the agreement of sentence-level annotations, agreement statistics are calculated using three different methods: The first two methods rely on Cohen's kappa, while the third analyzes the agreement rate of all labels in an essay. Refer to Appendix C for detailed information of these calculations.

Table 2 presents the results from the three agreement calculations, derived from the 15% sample of essays annotated by two annotators. On average, Approach 3 reflected the highest agreement, with values ranging from .75 to .79, indicating high agreement. Even though Approach 2 resulted in the lowest agreement values of the three approaches, ranging from .60 to .70, these values are nonetheless moderately high. This suggests that even the lower-bound estimate of annotator consistency supports that annotators were able to consistently apply the

argumentation labels to sentences within an essay. Across these three approaches, Prompt I exhibited the highest agreement, and the grade 6 prompts (A, B, and C) exhibited the lowest agreement.

Table 2. Annotator Consistency, by Prompt, using Three Agreement Metrices

Prompt ID	Grade	Agreement 1[a]	Agreement 2[b]	Agreement 3[c]
A	6	0.62	0.55	0.75
B	6	0.60	0.59	0.72
C	6	0.60	0.52	0.73
D	7	0.70	0.68	0.78
E	7	0.70	0.67	0.78
F	7	0.66	0.63	0.76
G	8	0.67	0.63	0.76
H	8	0.69	0.67	0.78
I	8	0.71	0.68	0.79
Average		0.66	0.62	0.76

[a]Cohen's kappa statistic computed from a single 7x7 contingency table; [b]Cohen's kappa statistic computed from seven 2x2 contingency tables, averaged across all annotation labels; [c]Exact agreement rate averaged across all essays.

The frequency of tag assignment by the annotators was as follows: Reasons, Evidence, Conclusion, Introduction, Main Idea, and, lastly, Transition. Table 3 presents Cohen's kappa statistics for each of the annotation labels, across all prompts. Generally, annotators exhibited high agreement for the labels of Introduction, Conclusion, and Main Idea, moderate agreement for Evidence and Reasons, and lower agreement for Opposing Position and Transitions. The Opposing Position agreement statistics were higher for those prompts associated with grade levels of seven and eight, which is when most students being to receive consistent instruction to include this element in their writing.

Table 3. Cohen Kappa Agreement by Argumentation Tag, Across Prompts)

Prompt ID	Grade	Main Idea	Intro.	Reas.	Evid.	Opp. Position	Trans.	Conc.
A	6	0.67	0.70	0.59	0.64	0.27	0.25	0.76
B	6	0.72	0.78	0.52	0.48	0.27	0.51	0.86
C	6	0.59	0.68	0.58	0.58	0.09	0.32	0.77
D	7	0.73	0.77	0.67	0.65	0.60	0.51	0.86
E	7	0.79	0.82	0.64	0.66	0.56	0.42	0.83
F	7	0.78	0.71	0.61	0.67	0.47	0.37	0.81
G	8	0.74	0.81	0.61	0.60	0.44	0.34	0.85
H	8	0.78	0.76	0.65	0.67	0.61	0.42	0.78
I	8	0.80	0.75	0.66	0.69	0.47	0.50	0.87

Note: Cohen's Kappa computed according to Approach # 2

Note: Main Idea = Main Idea; Intro. = Introduction; Reas. = Reasons; Evid. = Evidence; Opp. Position = Opposing Position; Trans. = Transition; Conc.= Conclusion

The hand-annotation process and results illustrate areas for refining the guidelines. For example, more clearly delineating the difference between *Evidence* and *Reasons* could help annotators better distinguish these tags. Additionally, more guidelines for detecting the subtleties of *Opposing Position* text may prove useful, as would providing a broader definition for the *Transitions* tag. Yet, the annotations, as they stand presently, were considered satisfactory to advance to the modeling stage. Before delving into the modeling process, we will first examine the selection criteria and guidelines pertaining to the conventions elements.

Conventions Elements

The third dimension of the rubric outlines five categories that reflect the foundations for the conventions rubric score: *Grammar Usage*, *Punctuation*, *Capitalization*, *Sentence Formation*, and *Spelling*. Smarter Balanced provides an additional rubric that offers detailed criteria for these criteria, delineating specific standards across various grade levels. Using these resources, the feedback tool was designed to detect the errors delineated in Figure 9, aimed at providing feedback tailored to students in grade 6 through 8. Work is currently underway to evaluate the performance of certain subcategories before their implementation, while others have been fully integrated. Specifically, ongoing evaluation is focused on the following: common noun capitalization, end-of-sentence punctuation, tense errors, and errors relation to determiners.

Conventions Category	Subcategory
Sentence Completion	Run-ons
	Fragments
Capitalization	Start of Sentence
	Common Nouns
	Personal Pronouns
	Proper Nouns
Punctuation	End-of-Sentence
	Apostrophes
	Commas
Spelling	Misspellings
Grammar Usage	Subject/Verb Agreement
	Tense
	Conjunctions - Incorrect
	Determiner
	Repeated Words
	Frequently Confused Words
	Adjectives/Adverbs

Figure 9. Conventions Categories and Subcategories

Labeling data for Conventions Elements

The dataset used to train the conventions model was a large synthetically created corpus of spell-correction data. More specifically, this dataset was derived by a process known as round-trip translation (Stahlberg and Shankar 2019), which takes a large set of grammatically correct sentences and pairs them with synthetically corrupted versions by translating to-and-from other languages. However, an exception to this process is the modeling for the *Sentence Formation* subcategories of fragments and run-ons. For this particular subcategory, additional data were needed because there are no publicly accessible datasets suitable for fine-tuning a pre-trained model for this particular task and because most available datasets do not reflect writing from middle-school-aged students. In particular, student writing often contains multiple conventions errors that can complicate the classification of sentences into run-ons and fragments. Using actual student data in training supports the model in addressing syntactic and morphological structures that may be unique to middle school writing.

To this end, eight editors at CAI participated in labeling approximately 10,000 sentences as either run-ons, fragments, or complete sentences. The sentences were sourced from the same data used for annotating argumentative elements. Editors were provided with guidelines outlining the rules for each sentence type. They assessed a set of 10 sentences at a time that were randomly chosen from the pool of sentences within essays written to various prompts. Inter-rater agreement was not evaluated.

These annotated essays, along with the identification of convention categories and subcategories, informed the subsequent modeling process. For the argumentation and sentence completion models, the training data was used to fine-tune models. For all other models, the defined list of conventions subcategories guided additional processing of the output from the pre-trained models.

AI Engine Training

The backend production engine consists of three final models responsible for receiving responses and returning labels. This section outlines the training process of the engine. Figure 9 provides an overview of the models at a high level. These models encompass a range of transformer-based models (Vaswani, 2017) alongside a rule-based component. They address argumentation elements, convention errors, and sentence completion, each with specific objectives. For example, the argumentation model was trained to predict sentence-level labels for responses, regardless of the essay prompt. The model for conventions-related errors is a generative large language model that produces a version of the text that is grammatically correct. In contrast, the sentence completion categorizes sentences as complete, run-on, or fragment, but doesn't recommend corrections. Figure 10 describes these models at a high level; more detailed information on modeling is provided in the following sub-sections.

	Argumentation Elements	Conventions Errors	Sentence Completion Errors
Model	XLNET LLM	T5 LLM and ERRANT Model	DeBERTa LLM
Objective	Create a generic argumentation model that is not prompt-specific.	Identify and classify grammar and spelling errors in student writing, and offer a specific solution to correct each error.	Predict one of three options for each sentence: A run-on, fragment, or correct.
Data Source	**Step 1**: Train five models using labeled data **Step 2**: Predict labels on unlabeled data/prompts **Step 3**: Train new model on predicted labels, and evaluate on original annotations	Grammar Error Correction dataset (Stahlberg and Kumar, 2021)	Sentences labeled from student essays sampled from the same source as the essays labeled for the argumentation model.

Figure 10. An Overview of the Models that Comprise the AI Engine

Modeling Argumentation Elements

A transformer-based model, XLNet (Yang et al. 2019), underwent fine-tuning to predict an argumentative label for every sentence within an essay. The model was fine-tuned to predict a special token labeled by the sentence annotation prepended to each sentence. Unlike many other transformer models, XLNet possesses the capability to handle essays of varying lengths, a crucial aspect of this project as it enables the model to consider the broader context of an essay when predicting specific argumentation elements for individual sentences.

The training approach for this model aimed to create a "generic model," ensuring that the annotations were suitable for all middle-school argumentative essays, irrespective of the prompt. The modeling training process consisted of three phases.

Initially, five models were trained, each on essays from a subset of the nine items containing human annotations. Subsequently, these models were utilized to generate synthetically-labeled annotations as follows: each of the five models predicted a label for every sentence of a sample of over 400,000 essays (from 21 prompts across grades 3-8) that were not human-annotated. These data came from data from the same state but from interim prompts (for informative and argumentative genres) and from summative prompts (for informative genres) across grades 3-8 and the same years the human annotated data were drawn from. The inclusion of another genre and grades was to create a large training dataset that was highly varied. The final label for each sentence was determined using the mode of the five model predicted labels, with any ties broken using the human annotation logical rules. These labels served as synthetic labels and were used to train a single annotation model.

Performance was then assessed using the original essays to train the initial five models. Results from this evaluation are presented in Table 4, by label and averaged across prompts. These results indicate that the model performs at, or above, human-human agreement levels for all but two elements. For these two elements, *Introduction* and *Transition*, the difference between human-human agreement and human-engine agreement amounts to a kappa value of .01.

Table 4. Human-Human Agreement and Human-Engine Agreement of Annotation Labels

Argumentative Elements	Human-Human Kappa	Human-Engine Kappa
Introduction	.76	.75
Controlling Idea	.74	.77
Evidence	.63	.68
Reasons	.62	.65
Opposing Position	.48	.53
Conclusion	.83	.83
Transition	.43	.42

To illustrate this agreement, refer to Figure 11, which displays human annotations on the left and the AI annotations on the right. The differences between the two are slight; and in cases of disagreement, they typically involve distinctions between *Evidence* and *Reasons*, mirroring disagreements observed among human raters.

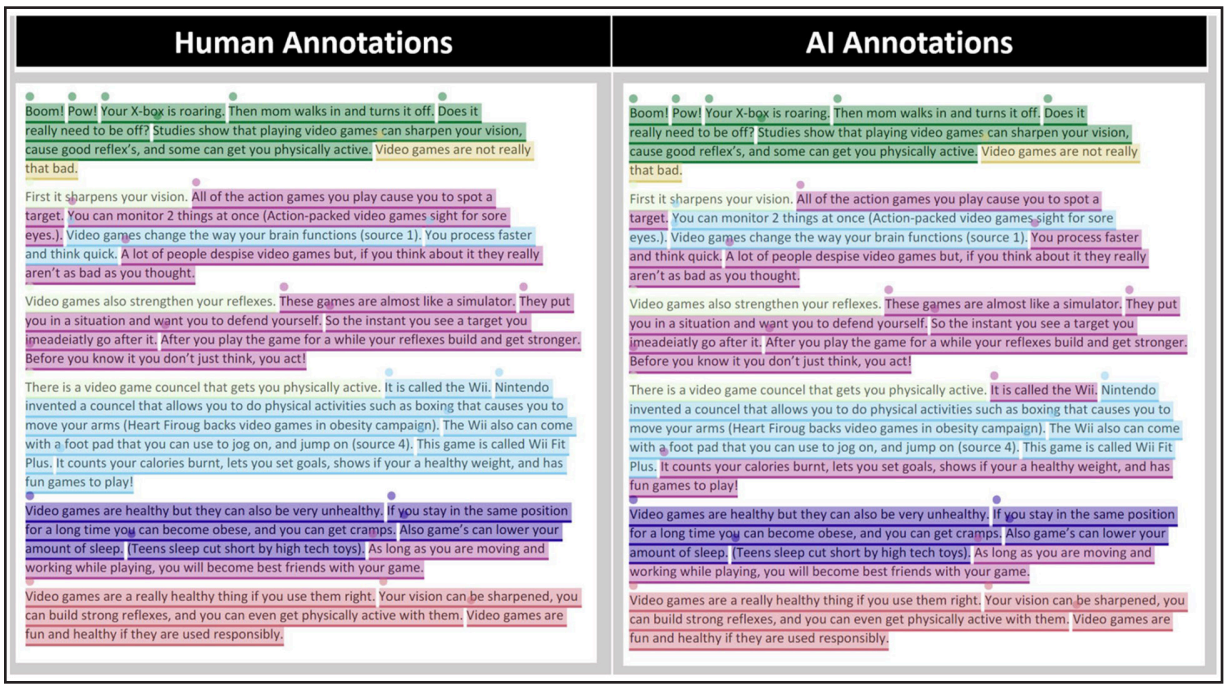

Figure 11. Machine and Human Annotations on a Single Essay

Modeling Conventions Errors

To detect most of the conventions errors, with the exception of sentence completion errors, two models were used jointly. First, a T5 model (Raffel et. al., 2019), a transformer model that possesses the same architecture used for translation tasks, was fine-tuned on a large corpus of synthetic data (Stahlberg and Shankar, 2019). In a typical translation task, the model receives text in one language as input and produces the same text translation into another language as output. In our case, for conventions, the input is a sentence with conventions errors, and the output is the corrected version of the sentence, where all conventions-related errors have been rectified. The original sentence and output sentence is then processed by a rule-based classifier known as ERRANT (Napoles et al., 2019) to classify the differences between the original and corrected versions.

The conventions model was benchmarked on the JFLEG dataset (Napoles et al., 2019), which consists of approximately 1400 development and 1400 test sentences, each with four possible acceptable corrections. By comparing n-grams and averaging across all corrected versions we obtain a GLEU score that can be used to evaluate model performance. The GLEU score for our conventions model (referred to as "Cambi"), in addition to the GLEU scores for some standard freely available software, has been presented in Table 5.

Table 5. A Comparison of GLEU Scores for a Variety of Spelling and Conventions Pipelines on the JFLEG Dataset

GLEU	Development	Human-Engine Kappa
Base	0.382	0.405
PySpell	0.425	0.462
LangTool	0.467	0.507
Cambi	0.559	0.611
Human	0.553	0.624

Modeling Sentence Formation

Finally, the DeBERTAa model (He et al., 2000) was fine-tuned to predict one of three classes for each sentence: complete, run-on, or fragment. Training data comprised a total of 6,360 sentences, with approximately 500 complete sentences, 800 run-ons, and 500 fragments. Kappa agreement and exact agreement were calculated using held-out validation data, as shown in the table below.

Table 6. Model Performance when Classifying Run-ons, Fragments, and Complete Sentences

Class (n)	Kappa	Exact Agreement
Complete sentences (1,068)	.71	90.4%
Run-on sentences (177)	.68	93.0%
Fragments (118)	.77	96.3%

As previously mentioned, these models were integrated to form a production engine responsible for receiving responses and providing all predicted labels corresponding to an essay. The following section outlines our process for crafting student-facing feedback that aligns with the predicted labels, aiding students in the editing process.

Crafting Feedback

This section focuses primarily on the feedback specific to the argumentation elements. On their own, the argumentation elements at the sentence level can provide valuable feedback to students regarding what is present and what is lacking in their writing; however, presenting these annotations without organized feedback may overwhelm students. Without structured guidance, students may struggle to understand their progress and determine their next steps. As such, an important guiding principle for this work is to provide feedback that highlights both progress and clear directions for next steps (Hattie, 2007). Another guiding principle is to craft feedback that supports both students and educators in such a way that is not disruptive to classroom practices. Importantly, the feedback within this tool is not intended to replace the role of a teacher. Rather the feedback should be aligned with teacher's instruction and serve as initial editing guidance, prior to the teacher engaging with the student and providing more substantive feedback. A final guiding principle is to provide feedback in such a way that employs language that is not accusatory of what a student may be missing, but rather remains constructive even in cases where the model may not be completely accurate.

As described in the first section of this paper, annotation-related feedback is divided into three sections: (1) *Organization*, (2) *Evidence* and *Reasons*, and (3) *Opposing Position* and *Rebuttal*. The figure below offers a condensed view of the feedback, emphasizing the essential steps in the process flow. Note that no feedback was crafted for the Transitions element because of the moderately-low agreement and the narrow definition of this element. Also, note that the feedback for *Rebuttal* is inferred when a *Reasons* annotation follows directly after an *Opposing Position* annotation.

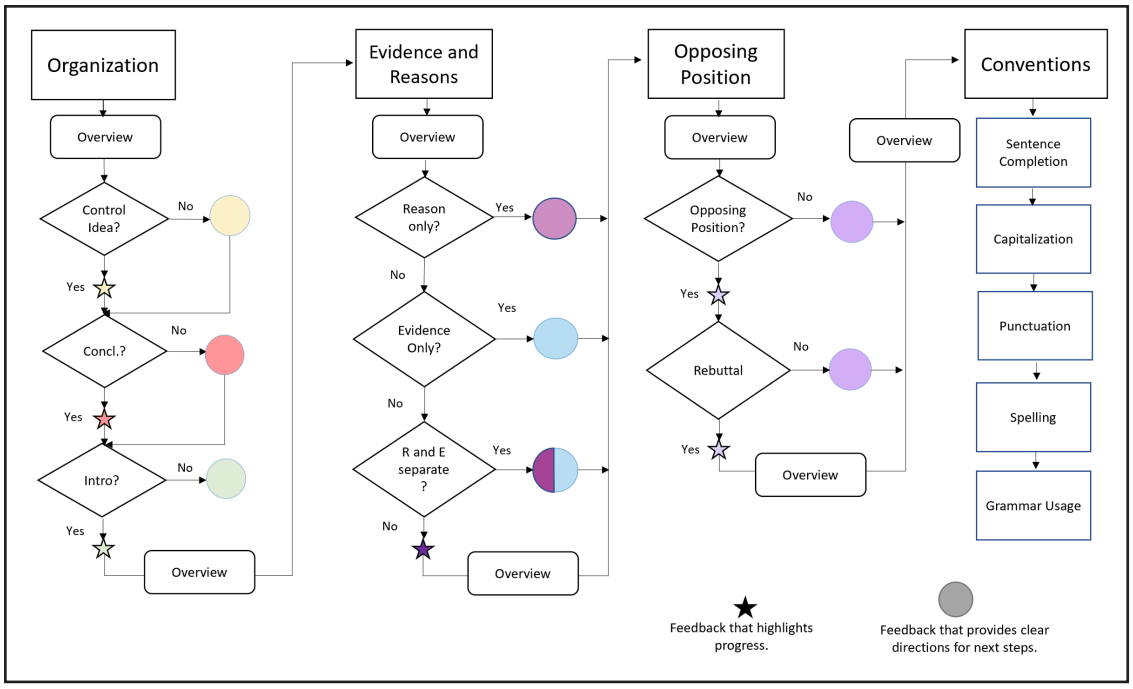

Figure 12. An Overview of the Feedback Flow within the Editing Tool

Teachers' Reactions to the Tool

The process of defining guidelines, labeling training data, building models, and crafting feedback represents key components of the project, culminating in the system featured at the outset of this paper. Throughout these stages, efforts were made to ensure alignment between the tool and state standards, as reflected in the rubric criteria. In this section, we delve into the alignment between the feedback provided by this tool and teachers' practices in the classroom, by conducting semi-structured interviews with teachers that allowed them to interact with the tool.

Twenty teachers participated in the interviews across two states. Fifteen educators expressed enthusiasm towards the use of the tool in their classrooms; four teachers were cautiously optimistic about the tool's potential, and one teacher was undecided in the use of the tool.

Through these interviews, teachers oftentimes reported that they viewed this tool as an assistant to help them address routine feedback at scale, which in turn could provide them with more time to work more closely with students to further develop their essay. One teacher stated, "If students used this tool, then I could actually talk to them about the argument they made, instead of the basics of an essay that they are missing."

Furthermore, teachers generally reported that the feedback provided by the tool was aligned with how they provided feedback in the classroom. One teacher noted, "It's really like the way I would teach. The questions that it guides them into like, *What else do you want to say about this?* and *What do you plan to say?* are all things that I would say as a teacher when trying to get a student to write an introduction." Another teacher noted similarities with exercises she assigns to help students with identifying different argumentative parts, "I like that it highlights the different parts for them. That's something that I would have them do in an assignment." Some discrepancies between teaching practices and the feedback from the tool were also noted. For example, one teacher reported that they discouraged first-person pronouns in an argumentative essay, whereas such feedback is currently not present in the tool.

Regarding which of their students would benefit most from the tool, teachers expressed a range of perspectives. Most thought that this tool could be useful to students who are struggling with writing. One teacher noted, "Some students can be overwhelmed by the idea of revision, and this tool helps them to narrow down what they are looking at. Some of my students may have no idea if they have written a main claim, or if they have written evidence. This tool could help them see where they may have to do something, and – perhaps – mimic later. This tool can also scaffold them in a way, if they know how to use it." Other teachers also recognized value in the tool's ability to remind students of the elements they oftentimes overlook. Many teachers also thought that this tool could foster independence and agency. As one teacher said, "I like that the tool allows the student to edit the essay if they think of something on their own. I like that because I don't want my students to be completely dependent on something else while writing. This tool is giving them a little extra push to go back and add things. It is not necessarily telling them what to say."

The results from these interviews suggested to us that "Write On with Cambi!" does align with teacher instruction in writing and can be useful to students. While not presented here, the teachers also offered recommendations for improving the tool such as adding text to speech, providing instructional information on each feedback element to better support students as they revise their essays, and improving the UI/UX design in ways the reduce the reading burden and cue students to next steps.

Discussion and Conclusion

This paper introduces “Write On with Cambi!”, an AI-supported editing tool, and outlines its internal components designed to ensure high-quality feedback to students.

As with most AI-based methods, actual student responses and human ratings serve as the basis for modeling. These ratings – or annotations in this case – also need to be thoughtfully created in ways that represent the core constructs we are intending to measure. With regard to the argumentation annotations, we found that the human annotators consistently apply sentence-level labels to detect key elements in argumentative essays. This success can be attributed, in part, to the quality of the annotation guidelines and the expertise of the hand-scorers. The features related to conventions also seem to function effectively on our held-out datasets. We will continue to monitor their performance as students interact with the tool and explore opportunities to expand the range of subcategories addressing conventions-related feedback.

Based on the evaluation metrics examined, the machine learning models predict the various finer-grained elements of student writing with acceptable levels of accuracy, The evidence for this is particularly strong for the model predicting argumentation elements. This result is very promising because it indicates that fine-grained feedback, grounded in rubrics, high-quality hand-scoring, and transformer-based modelling, is possible to provide to students.

The guiding principles for crafting feedback were in the spirit of supporting both students and educators in a way that would be perceived as helpful without interfering with a teacher’s classroom practices. The positive feedback from teachers suggests that we are on the right track with this effort. As one teacher described the tool, “I want to use it! I think it’s brilliant. I like this because there’s still a lot of ownership on the kids and on me -- as well as human interaction, which is really important when it comes to writing, because we’re talking about the thinking process. I think this is a great example of how AI could be a very useful tool without taking over.” To this end – developing an AI tool that does not overreach – it is important to attend to the various recommendations by teachers, which include finding a way to provide feedback to a student that is more concise, and allowing for flexibility to align with localized language and instruction.

Ongoing work will be conducted to continue to understand how teachers and students are using the tool. The long-term benefits of this will be, of course, that this editing support is something that students can learn from, and – perhaps, as one teacher noted – mimic later.

Appendix A
ICCR Version of the Rubric

Grades 6–8 • Argumentation Text-based Writing Rubric
(Score points within each domain include most of the characteristics below.)

Score	Purpose, Focus, and Organization (4-point Rubric)	Evidence and Elaboration (4-point Rubric)	Conventions of Standard English (2-point Rubric begins at score point 2)
4	The response is fully sustained and consistently focused within the purpose, audience, and task; and it has a clear claim and effective organizational structure creating coherence and completeness. The response includes most of the following: • Strongly maintained claim with little or no loosely relatedmaterial • Clearly addressed alternate or opposing claims* • Skillful use of a variety of transitional strategies to clarify therelationships between and among ideas • Logical progression of ideas from beginning to end with asatisfying introduction and conclusion • Appropriate style and tone established and maintained	The response provides thorough, convincing, and credible support, citing evidence for the writer's claim that includes the effective use of sources, facts, and details. The response includes most of the following: • Smoothly integrated, thorough, and relevant evidence, including precise references to sources • Effective use of a variety of elaborative techniques to support the claim, demonstratingan understanding of the topic and text • Clear and effective expression of ideas, using precise language • Academic and domain-specific vocabulary clearlyappropriate for the audience and purpose • Varied sentence structure, demonstrating language facility	
3	The response is adequately sustained and generally focused within the purpose, audience, and task; and it has a clear claim and evident organizational structure with a sense of completeness. The response includes most of the following: • Maintained claim, though some loosely related material maybe present • Alternate or opposing claims included but may not becompletely addressed* • Adequate use of a variety of transitional strategies toclarify the relationships between and among ideas • Adequate progression of ideas from beginning to end with asufficient introduction and conclusion • Appropriate style and tone established	The response provides adequate support, citing evidence for the writer's claim that includes the use of sources, facts, and details. The response includes most of the following: • Generally integrated and relevant evidence fromsources, though references may be general orimprecise • Adequate use of some elaborative techniques • Adequate expression of ideas, employing a mix ofprecise and general language • Domain-specific vocabulary generally appropriatefor the audience and purpose • Some variation in sentence structure	
2	The response is somewhat sustained within the purpose, audience, and task but may include loosely related or extraneous material; and it may have a claim with an inconsistent organizational structure. The response may include the following: • Focused claim but insufficiently sustained or unclear • Insufficiently addressed alternate or opposing claims* • Inconsistent use of transitional strategies with little variety • Uneven progression of ideas from beginning to end with aninadequate introduction or conclusion	The response provides uneven, cursory support/evidence for the writer's claim that includes partial use of sources, facts, and details. The response may include the following: • Weakly integrated evidence from sources; erratic or irrelevant references or citations • Repetitive or ineffective use of elaborative techniques • Imprecise or simplistic expression of ideas • Some use of inappropriate domain-specific vocabulary • Most sentences limited to simple constructions	The response demonstrates an adequate command of basic conventions. The response may include the following: • Some minor errors in usage but no patterns of errors • Adequate use of punctuation, capitalization, sentence formation, and spelling
1	The response is related to the topic but may demonstrate little or no awareness of the purpose, audience, and task; and it may have no discernible claim and little or no discernible organizational structure. The response may include the following: • Absent, confusing, or ambiguous claim • Missing alternate or opposing claims* • Few or no transitional strategies • Frequent extraneous ideas that impede understanding • Too brief to demonstrate knowledge of focus or organization	The response provides minimal support/evidence for the writer's claim, including little if any use of sources, facts, and details. The response may include the following: • Minimal, absent, erroneous, or irrelevantevidence or citations from the source material • Expression of ideas that is vague, unclear, orconfusing • Limited and often inappropriate language ordomain-specific vocabulary • Sentences limited to simple constructions	The response demonstrates a partial command of basic conventions. The response may include the following: • Various errors in usage • Inconsistent use of correct punctuation, capitalization, sentence formation, and spelling
0			The response demonstrates a lack of command of conventions, with frequent and severe errors often obscuring meaning.

Appendix B
Annotation Guidelines

Annotation Rules

1. Apply a single tag per sentence using the chart below.
2. If more than one tag applies, use the following ordered list and apply the tag that appears highest in the hierarchy:
   a) Controlling Idea
   b) Opposing Position
   c) Evidence
   d) Elaboration
   e) Transitions
3. Tag sentences by giving student benefit of doubt with regards to the accuracy of citation or analysis

Annotation Tag	Annotation Definition	What should be highlighted	What should NOT be highlighted
Introduction	Context to explain the issue	• Plan for argument, such as listing out subtopics (organizational outline) • Use of rhetorical devices to help reader feel like this issue is important • Attention grabbing devices (which may include citation/quotation of sources) • Stage setting sentences	The controlling idea in the introduction is tagged as controlling idea (see below)
Controlling Idea	Argumentative-Overarching statement about the debatable issue.	• The stand that the author is taking on an issue or topic • Taking center of the road position • Flip flops (highlight both claims)	All of the sub-topics of Controlling Idea (this is elaboration or introduction).
Evidence	Integration of evidence, including data, statistics or similar study results, and quotations to support controlling idea or sub-topics	• Citation/quotation of sources • Data from sources • Paraphrasing authors as a citation.	
Elaboration	Explanations, elaborations, and interpretations of evidence and Controlling Idea	• General supports of claim • Reasons • Sub-claims • Rhetorical questions or other rhetorical devices (e.g., allusions, appeal to logos, repetition, parallelism, fragments for effect) to enhance arguments) • Relevant commentary and examples • Definitions for related terms • Rebuttal of opposing position in support of controlling idea	Extraneous information, such as an entire paragraph that does not seem relevant to the controlling topic or a sentence that jumps out as being off-topic.
Opposing Position	Acknowledge opposing side of argument	• Acknowledgement of the opposing position (not in intro or conclusion) • Sentences that explore or explain the opposing position	
Conclusion	Summation of Controlling Idea	• Summary of sub-topics, evidence, and elaboration • Tug-at-heartstring moments or call to action • Use of rhetorical devices to garner support (which may include citation/quotation of sources)	Introduction of new ideas
Transitions	Signposts to help guide the reader through the development of the essay	• Sentences at the end of a paragraph signaling what is coming next • Sentences at the beginning of the paragraph reiterating a sub-claim detailed in the introduction • Sentences with no new information intending to create coherence or structure. • Sentences that re-state introductory organizational phrases to guide the sub ideas of each paragraph	Sentences that use transitional clauses but contain evidence, elaboration, or opposing positions. These should be tagged by using the hierarchy listed at the top of the table
None of the Above	Off-topic and extraneous information		

Appendix C
Description of Agreement Statistics

Cohen's kappa requires two pieces of information: The proportion of sentences where the two raters agreed with one another, and the expected agreement of the sentences based on chance (calculated based on the independent probabilities of the ratings by the two annotators). The distribution of ratings for the first and second annotator are computed using a contingency table. The rows and columns can be referred to as the rating variables – and herein lies the difference between the two approaches. The first approach (Approach #1) considers each of the seven annotation labels as rating variables; all sentences from all essays are included in the same 7x7 contingency table. The second approach (Approach #2) computes a 2x2 continency table for each of the seven annotation labels. For each sentence, the presence of an annotation label is marked by a value of '1' and the absence is marked by a value of '0'. In this second approach, a single, aggregated kappa value is computed by averaging all of the kappa values together. Both aggregated and disaggregated values are presented in this paper. Finally, the third approach (Approach #3) is computed for each essay in adherence to the following: For each sentence, if the annotation label matches for the two annotators, mark this agreement as a '1' and otherwise, mark the sentence as a disagreement, '0'. Then, to arrive at an agreement rate for each essay, take the average of all values. A single agreement statistic is computed by averaging across all essays.

References

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. Assessing Writing, 54, 100667.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of educational research, 77(1), 81-112.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA.

Kaggle. (2022, May). Feedback Prize - Predicting Effective Arguments. Retrieved from https://www.kaggle.com/competitions/feedback-prize-effectiveness/overview/

Lottridge, S., Burkhardt, A., Dwyer, K., & Woolf, S. (2023, April). RALD- and rubric-based assertions and annotations: Defining and automating feedback to students. Paper presented at the National Council on Measurement in Education (NCME) Annual Meeting, Chicago, IL.

Ormerod, C., Burkhardt, A., Young M., and Lottridge. (2023). Argumentation element annotation modeling using XLNet. [Manuscript In preparation].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arxiv:1910.10683

Smarter Balanced. (2022, August). Argumentative Performance Task Writing Rubric (Grades 6 – 11). Retrieved from https://portal.smarterbalanced.org/library/en/performance-task-writing-rubric-argumentative.pdf

Stab, C., & Gurevych, I. (2014, August). Annotating argument components and relations in persuasive essays. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers (pp. 1501-1510).

Stahlberg, F., & Kumar, S. (2021). Synthetic Data Generation for Grammatical Error Correction with Tagged Corruption Models. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 37–47). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2021.bea-1.4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

West Virginia Department of Education. (n.d.). ELA Text-based Writing Rubrics Grades 6–8: Argumentation. Retrieved from https://wv.portal.cambiumast.com/-/media/project/client-portals/west-virginia/pdf/2022/wv-ela-text-based-writing-rubrics---grades-3-8.pdf

Yang Z., Dai., Z., Yang, Y., Carbonell, J., Salakhutdinov R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arxiv:1906.08237.