Don't Score too Early!
Evaluating Argument Mining Models on Incomplete Essays

Nils-Jonathan Schaller[1], Yuning Ding[2], Thorben Jansen[1], Andrea Horbach[1]
[1]Leibniz Institute for Science and Mathematics Education at the University of Kiel, Germany
[2] FernUniversität in Hagen, Germany
schaller@leibniz-ipn.de

Abstract

Students' argumentative writing benefits from receiving automated feedback, particularly throughout the writing process. Argument Mining (AM) technology shows promise for delivering automated feedback on argumentative structures; however, existing systems are frequently trained on completed essays. Although they provide rich context information, concerns have been raised about their usefulness for offering writing support on incomplete texts during the writing process. This study evaluates the robustness of AM algorithms on artificially fragmented learner texts from two large-scale corpora of secondary school essays: the German DARIUS corpus and the English PERSUADE corpus. Our analysis reveales that token-level sequence-tagging methods, while highly effective on complete essays, suffer significantly when the context is limited or misleading. Conversely, sentence-level classifiers maintain relative stability under such conditions. We show that deliberately training AM models on fragmented input substantially mitigates these context-related weaknesses, enabling AM systems to better support dynamic educational writing scenarios.

1 Introduction

Providing in-process and constructive feedback is integral to fostering argumentative writing skills in educational settings. (Argument Mining) AM has emerged as a promising approach for analyzing and evaluating the structure and quality of argumentative learner essays. However, the majority of AM systems operate on complete texts and their applicability in dynamic writing scenarios with incomplete drafts is unknown. A *conclusion*, for example, might only be recognized as such because it appears at the end of an essay and might not be recognized if the learner starts with the concluding statement and requests feedback early on.

To assess the severity of this problem, we investigate the robustness of existing AM algorithms when applied to incomplete essay texts. We do so by emulating work-in-progress texts by applying artificial changes and perturbations to two datasets used previously for educational AM: the German DARIUS (Digital Argumentation Instruction for Science) dataset (Schaller et al., 2024b), containing about 4,500 texts on socio-scientific issues, and the English PERSUADE dataset (Crossley et al., 2024), with over 25,000 essays written by US secondary school students on various topics.

These benchmark datasets are probed with two kinds of AM classifiers: a sequence tagger that assigns a label to each token within an essay and a sentence classification approach that labels individual sentences without context. Although sequence tagging exploits contextual information and thus seems suitable for complete essays, sentence classification might have an advantage when only little or misleading context is available. To pave the way for feedback algorithms that can already provide support during the writing process, we explore ways to train a more robust classifier by applying similar perturbations to the training data.

Our paper makes the following contributions:

- We provide benchmark datasets of essays that we corrupted in several more or less realistic ways, which we obtained from the existing PERSUADE and DARIUS datasets to foster the development of robust AM methods.

- We conduct experiments on these datasets to highlight the detrimental effect of incomplete input in educational AM (up to 22 percentage points in the F1 score compared to full texts).

- We train baseline classifiers on similarly corrupted data that can reduce this performance drop to less than half.

All data is made available: https://github.com/darius-ipn/dontscoretooearly

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 345–355
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

We review related work regarding three aspects. First, we discuss existing datasets for educational AM and select those suitable for our experiments. We then examine machine learning approaches in AM, focusing on the distinction between sequence tagging and sentence classification methods. Finally, we discuss other studies using perturbed essay texts in automated scoring scenarios.

2.1 Educational Argument Mining Datasets

We review datasets according to their suitability for our experiments considering their size, language, and the argumentative units annotated in the data.

Stab and Gurevych (2017) developed an argument detection system for English persuasive essays, achieving substantial inter-annotator agreement (Krippendorff's α_U of 0.77). Their data consist of 402 English essays annotated for *major claim*, *claim*, and *premise*, as well as their argumentative relationships (*support*/*attack*).

Wambsganss et al. (2020) built a feedback system for the argumentation structure of German students' essays. For that purpose, they collected 1,000 peer-reviews from a business innovation course in which students evaluated each other's business models. The corpus was annotated by three native German-speakers for argumentative components (*claim* and *premise*) and their relationships.

The PERSUADE dataset of Crossley et al. (2022) consists of over 25,000 argumentative essays written by secondary school students in the US. Each essay was annotated for seven distinct argumentative components. The dataset was expanded (Crossley et al., 2024) with effectiveness scores per unit and holistic scores for the overall essay quality. In our previous work (Schaller et al., 2024b) we compiled the DARIUS corpus of 4,589 argumentative essays written by secondary school students in Germany. The corpus consists of two writing prompts on socio-scientific topics. The corpus features detailed annotations of argumentative elements, including *content zone*, *major claim*, *position*, and *warrant*.

Stahl et al. (2024) presented a German corpus of 1,320 school student essays annotated for argumentative structure and quality. Their four-level annotation scheme achieved high agreement (α = 0.74-0.89). Their analysis revealed significant correlations between structural elements and essay quality. The corpus provides another reference point alongside DARIUS for student writing in German secondary schools.

Velentzas et al. (2024) presented KUPA-KEYS, a dataset of keystroke logs from 1,006 participants' English essays, including both L1 and L2 writers. Each essay was evaluated on the CEFR scale by three human assessors and an automated system. The dataset captures detailed keystroke patterns, pauses, and revisions during writing tasks, with the analysis showing moderate correlations between keystroke patterns and writing proficiency.

While keystroke logging would be ideal for studying incomplete texts as it captures the authentic writing process, existing keystroke datasets such as KUPA-KEYS lack the specific argumentative annotations needed for our work.

We decided to use the DARIUS and PERSUADE datasets for our experiments because they offer several key advantages: First, they cover different languages (German and English), allowing us to verify whether our findings hold across languages. Second, they are very similar in their annotation schemes, allowing us to focus on sequences instead of whole documents. Third, with over 4,500 and 25,000 essays respectively, they provide sufficient data to train robust machine learning models and conduct comprehensive robustness evaluations

2.2 Machine Learning in Argumentation Mining

AM consists of two subtasks: the detection of argument units and their classification as a certain type of argument, e.g., a claim or a conclusion.

Approaches to Argument Unit Detection There are two main strategies for detecting argumentative units, although variations are possible: Some use sentence classification, treating entire sentences as argumentative units. Alternatively, sequence tagging works at the token level to identify more flexible argument boundaries. We further review both approaches and hybrid methods.

Sentence Classification Approaches An easy (but not necessarily optimal) method for the selection of units is sentence classification, thus omitting explicit argument detection and classifying individual sentences in a text as belonging to a certain type of argument (or as being non-argumentative). Wambsganss et al. (2020) built a feedback system for the argumentation structure of students' German texts based on a sentence-level multiclass classification task. They developed an SVM-based system for claim/premise identification (65.4% accuracy) and relationship classification (72.1% accuracy). Their later ArgueTutor system (Wambsganß et al., 2021) improved performance using BERT (F1 = .73). Similarly, in their fairness investigation, Schaller et al. (2024a) employed sentence classification approaches among others, comparing a supervised SVM, a BERT-based classifier, and zero-shot GPT-4 on the DARIUS corpus of students' German essays.

Sequence Tagging Approaches At the other end of the spectrum, many researchers have employed sequence tagging on tokens to allow for more flexible argument boundaries. This approach has gained significant attention in the field, particularly with the rise of transformer-based models. In our fairness study, (Schaller et al., 2024a) we additionally employed sequence tagging approaches, finding that a task-specific fine-tuned BERT model consistently outperformed other approaches, including more powerful decoder-based language models such as GPT-4 when used in a zero-shot setting.

Stahl et al. (2024) trained sequence labeling models based on mDeBERTaV3-adapter, which achieved a F1 scores up to .68 for discourse functions.

We previously investigated AM using the English PERSUADE and MEWS and the German DARIUS datasets (Ding et al., 2024) . Our sequence tagging used a Longformer-based model, achieving an F1 score of .66 on English essays and providing important baseline performances on complete essays. The analysis revealed that educational context differences impacted performance more than language differences did.

Comparative and Hybrid Approaches Several researchers have compared these approaches or examined the advantages of hybrid approaches.

Trautmann et al. (2020) explicitly examined sentence versus token classification for argument recognition on annotated Common Crawl data (IAA $\alpha_{u_{nom}}$ = .61). Their experiments with various BERT and FLAIR models showed that a BERT_LARGE sentence classifier was only outperformed by the BERT_LARGE token classifier when combined with a CRF model, demonstrating that sentence classification can achieve comparable results to token classification approaches.

Stab and Gurevych (2017) showed the advantages of combining both sequence labeling and classification for AM. They first used a CRF for sequence labeling to identify argument boundaries in the text and then used SVM classification to determine each argument's type and relationships. This combined approach significantly outperformed using either method alone, achieving an F1 score of .86 compared to .79 for individual classification and .64 for baseline approaches. These studies highlight that the optimal approach for complete texts may depend on specific tasks, domains, and available data, and that one approach may not always be superior.

2.3 Influence of Rearranged Sequences

As our work investigates the robustness of AM on incomplete texts, we review previous work on educational scoring that examined model behavior with incomplete texts or nonstandard text order.

Farag et al. (2018) tested AES robustness against shuffled sentences. Their LSTM model performed well on regular essays but declined with shuffled texts. They addressed this by combining models trained on both regular and permuted essays, maintaining scoring performance while detecting shuffled texts. This highlights neural models' reliance on expected sequence orders, informing our work on incomplete texts.

In a previous paper (Ding et al., 2023), we used sequence tagging with fine-tuned RoBERTa to identify EFL email segments. When tested with scrambled segments, performance dropped strongly (F1 = .89 to F1 = .60), but the model was able to adapt when trained on scrambled data (F1 =.85). On the basis of this work, we anticipate that sentence classification models will outperform sequence tagging models due to lower dependence on position and context.

3 Data

This section presents the two datasets used in our evaluations, DARIUS and PERSUADE, and discusses our decision regarding which annotations to include. Table 1 gives an overview of the key statistics of the datasets.

DARIUS The DARIUS corpus (Schaller et al., 2024b) contains 4,589 argumentative texts by 1,839 secondary school students in 33 German schools. The task consists of two writing prompts on socio-scientific topics: energy and automotive (gener-

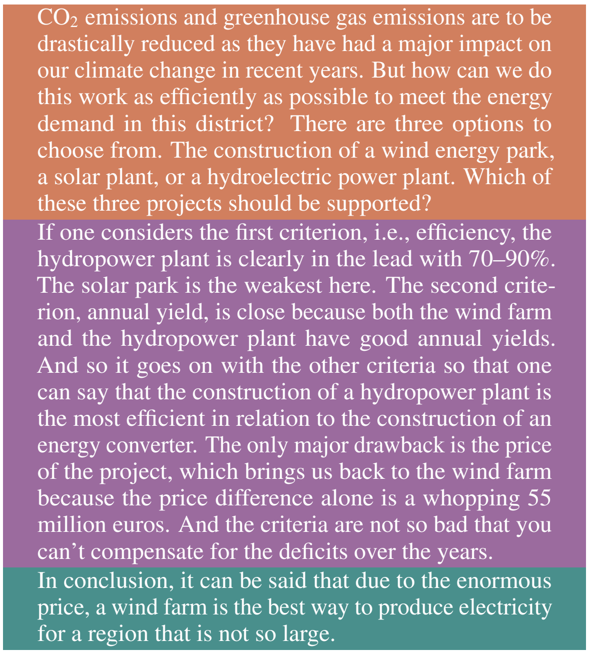
	DARIUS	PERSUADE
Language	German	English
Essay genre	argumentative	argumentative
Writing prompts	2	15
Grade	9-13	6-12
# Essays	4,675	25,996
avg. Word count	150	399

Table 1: Key statistics for DARIUS and PERSUADE

ating 2,307 and 2,282 essays, respectively), with students completing both a draft and a revision for one prompt and a single essay for the other. The corpus features detailed annotations of argumentative elements, including *content zone*, *major claim*, *position*, and of argumentative units. The average essay has 150 words.

For our experiments, we focused on *content zone* annotations. *Content zone* describes the macro structure of an essay: *introduction*, *main part*, and *conclusion* (see Fig. 1). Not all essays contain all three sequences, as students may have returned unfinished texts or skipped introductions, etc. Each sequence consists of one or more complete sentences and must end with a sentence-final punctuation mark. In contrast to the other annotations in DARIUS, their sequences are not based on sentences but on spans of multiple sentences, similar to the annotations of PERSUADE.

PERSUADE The PERSUADE dataset (Crossley et al., 2022) consists of over 25,000 argumentative essays written by secondary school students from grades six through twelve in the US. The essays are based on 15 prompts, eight independent and seven source-based. Each essay was annotated for seven distinct argumentative components: (see Fig. 2 and Fig. A.6 in Appendix A) *lead*, *position*, *claim*, *counterclaim*, *rebuttal*, *evidence*, and *concluding statement*. The inter-rater agreement achieved an F1 score of .73. The dataset was expanded (Crossley et al., 2024) to include effectiveness scores for individual discourse elements and holistic essay quality scores. The average essay has 399 words. Sequences can span any length from a single phrase to multiple sentences and do not always align with sentence boundaries.

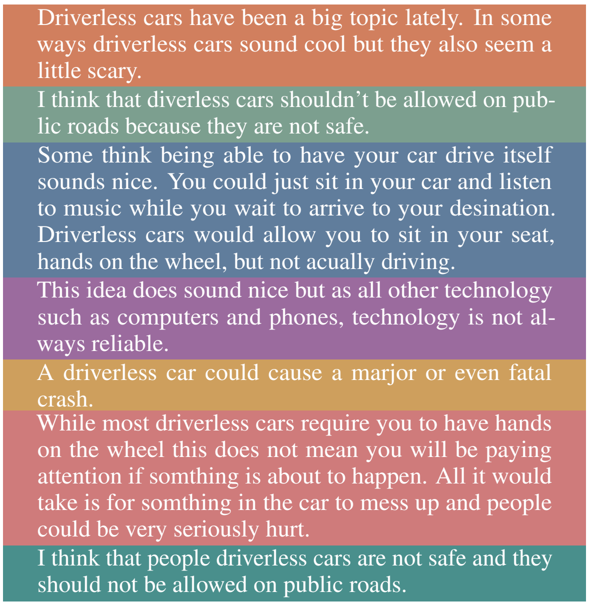


CO_2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years. But how can we do this work as efficiently as possible to meet the energy demand in this district? There are three options to choose from. The construction of a wind energy park, a solar plant, or a hydroelectric power plant. Which of these three projects should be supported?

If one considers the first criterion, i.e., efficiency, the hydropower plant is clearly in the lead with 70–90%. The solar park is the weakest here. The second criterion, annual yield, is close because both the wind farm and the hydropower plant have good annual yields. And so it goes on with the other criteria so that one can say that the construction of a hydropower plant is the most efficient in relation to the construction of an energy converter. The only major drawback is the price of the project, which brings us back to the wind farm because the price difference alone is a whopping 55 million euros. And the criteria are not so bad that you can't compensate for the deficits over the years.

In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large.

Figure 1: Example essay from the DARIUS dataset (translation by the authors). From top to bottom: *introduction*, *main part*, *conclusion*.



Driverless cars have been a big topic lately. In some ways driverless cars sound cool but they also seem a little scary.

I think that diverless cars shouldn't be allowed on public roads because they are not safe.

Some think being able to have your car drive itself sounds nice. You could just sit in your car and listen to music while you wait to arrive to your desination. Driverless cars would allow you to sit in your seat, hands on the wheel, but not acually driving.

This idea does sound nice but as all other technology such as computers and phones, technology is not always reliable.

A driverless car could cause a marjor or even fatal crash.

While most driverless cars require you to have hands on the wheel this does not mean you will be paying attention if somthing is about to happen. All it would take is for somthing in the car to mess up and people could be very seriously hurt.

I think that people driverless cars are not safe and they should not be allowed on public roads.

Figure 2: Essay 65 from the PERSUADE corpus. From top to bottom: *Lead*, *Position*, *Claim*, *Counterclaim*, *Rebuttal*, *Evidence*, *Concluding Statement*.

4 Benchmark Datasets

In order to gauge classifier performance on incomplete essay data, we simulate essays in the process of being written by making different changes and perturbations to the original essay data.

Across all benchmark variants, we retain the original gold-standard labels from the uncorrupted texts, i.e., we assume that the label of a unit does not change depending on the context or lack thereof. In doing so, we want to simulate a process where students write what they intend to write but not necessarily in the right order, and the justification for a label is that the unit has a certain function in the complete text even if that is not yet obvious to a teacher or annotator in the partial essay during the writing process. While this may lead to cases where labels are not recoverable from the modified input alone (e.g., conclusions appearing mid-essay in Shuffled texts), our goal is to test model robustness in incomplete or misleading contexts, reflecting educational scenarios where students submit partial drafts.

We include the following variants:

Full text: The original essays are used as a baseline.

First_X: This simulates an incremental, top-down writing process where a learner receives feedback after having completed at least 25%, 50%, or 75% of the whole text. We clip DARIUS essays after 25% of sentences and PERSUADE essays after 25% of sequences, as DARIUS annotations span at least one sentence, while PERSUADE annotations can be phrase-level annotations.

Last_X: Similar to the former scenario, this simulates the (admittedly less likely) process where only the last 25%, 50%, or 75% of a text is scored.

Sentence: Every single sentence from the test data as its own document, i.e., sentences are labeled completely without context.

Shuffled: For this condition, sequences within an essay are randomly shuffled to appear in misleading contexts, with DARIUS shuffling at the sentence and PERSUADE at the sequence level.

Table A.7 in Appendix A provides examples for all variants of an essay. Note that we do not expect all of these variants to occur (frequently) in real-life data. We rather aimed to cover the whole spectrum to also assess worst-case scenarios. We kept the datasets comparable whenever possible, i.e., **full text**, **sentence**, and **shuffle** contain exactly the same material, just in a different order or different document size, while **first_X** and **last_X**, for obvious reasons, contain less material. Similarly, also the label distribution changes for these benchmarks, e.g., essays with the last part missing, tend to contain less conclusion material. Table A.5 in Appendix A shows an overview of the dataset sizes per benchmark.

5 Experiments

This section evaluates how AM approaches perform on incomplete texts that simulate intermediate products within the writing process.

We compare sequence tagging and sentence classification. The models are tested on both complete essays and simulated incomplete drafts to assess their ability to provide feedback on unfinished texts.

5.1 Experiment 1: Baseline Performance on Benchmark Datasets

Data splits The DARIUS corpus consists of 4,581 essays, which we divided into 3,672 essays for training and 909 essays for testing, following the setup from Schaller et al. (2024a). From the training set, we reserved 20% as a development set to determine the optimal number of training epochs, while using the remaining 80% for model training.

The training set of the PERSUADE (1.0) corpus contains 15,594 essays, which were released in a Kaggle competition[1]. We use them all for training. For testing, we use 1,560 random essays from the PERSUADE (2.0) test set[2].

Classifiers For the comparison of sequence tagging and sentence classification, we use transformer models with different classification heads while keeping the base architecture consistent within each language. For DARIUS, we use a BERT (Devlin et al., 2019) model, bert-base-german-cased, both for the sequence tagging and the sentence classification model. It has a sequence length of 512 tokens, which is adequate for the average DARIUS essay. Longer texts were truncated to 512 tokens. The model was fine-tuned using the default hyperparameters from the Hugging Face Transformers library. We trained for four epochs, which proved sufficient for convergence on our validation set.

As almost one-third of the English essays in PERSUADE contain more than 512 tokens, we use a pretrained Longformer model (Beltagy et al., 2020) for token classification, with a maximal training length of 1,024 tokens, to train a sequence tagging pipeline (Ding et al., 2022) for the prediction of different argumentative elements in the PERSUADE essays. For sentence classification, we use

[1] https://www.kaggle.com/competitions/feedback-prize-2021

[2] https://github.com/scrosseye/persuade_corpus_2.0

the pretrained BERT model (Devlin et al., 2019) for sequence classification.

Evaluation Measures We use F1 scores to evaluate and compare sequence tagging and sentence classification approaches across different benchmark datasets. For both datasets, we treat each argumentative element type (e.g., *introduction*, *main*, and *conclusion* in DARIUS; *lead*, *position*, *claim*, etc. in PERSUADE) as a separate class in a multi-class classification setting.

For sequence tagging, we compute F1 scores directly on the token-level predictions, where each token receives one label. When comparing this to sentence classification approaches, we also derive sentence-level metrics from our token classifiers through majority voting (assigning the most frequent token label as the sentence label).

5.2 Experimental Results

Results for DARIUS Table 2 shows the performance of the token and sentence classifier for the various benchmark datasets.

On complete essays, the token classifier demonstrated high performance, with an overall F1 score of .93 on the sentence test set and .92 on the token test set, showing particular strength in identifying the *introduction* (.91/.88) and *main* sections (.96/.95), although it performs weakly on the *conclusion* (.73/.71). This will be the baseline for the other benchmark datasets. As the results for both datasets are very similar, we will further discuss only the sentence test set.

With regard to the decontextualized datasets, it can be seen that the performance decreases. When tested on the first 25%/50%/75% of sentences, relatively stable F1 scores are observed: *overall* (.93-.96), *introduction* (.91-.92), and *main* (.94-.97). The drastic drop in *conclusion* performance (.0-.56) stems directly from the near-complete absence of conclusion sentences in these portions. In the first 50% of essays, there are virtually no conclusion sentences (30 instances compared to 13,679 for the full texts), explaining the .0 F1 score. Most conclusion sentences appear only in the last 25% of essays, as confirmed by the higher support numbers in the Last_25 and Last_50 datasets.

Testing on the **Last_75** reveals the reversed situation, reflecting the lack of *introductions* while maintaining the same number of *conclusions* as in the full essays. The F1 scores for the *overall*, *main*, and *conclusion* remained stable, although the *introduction* F1 score decreased substantially to .76.

The analysis of the **Last_25** demonstrated a substantial drop in the *overall* F1 score (.72) and the *conclusion* (.32) - a seemingly unexpected result given that all *conclusion* samples remain present in this set. This decline occurs because conclusion sections now appear at the beginning of these truncated texts. The model struggles to recognize conclusions when they are artificially repositioned, indicating that it relies on positional context.

An even worse picture emerges from the analysis of the **Sentence** and **Shuffled** benchmark data. Both include all samples of each annotation but an extreme drop can be seen in the *conclusion*. The token classifier, in particular, is not able to predict the *conclusion*, if given only a sentence example of it (.0). The *introduction* also drops to an F1 score of .70. In the shuffled condition, the *introduction* performance declines further to .31.

These results suggest that, although the model learns the typical structure of complete essays, it struggles to apply this knowledge to incomplete texts.

To further investigate this assumption, we also trained a model for sentence classification, see Table 2. Compared to the token classifier, this model has a slightly lower overall F1 score of .91, .86 for the *introduction* and a slightly higher F1 score of .96 for *main* when tested on the **Full text**. It also performs lower on the *conclusion*, with a score of .60. But compared to the performance of the token classifier on sentences (.0), a much better performance was observed here, especially on the *conclusion*. This might indicate that the token classifier heavily relies on the context of each class, whereas the sentence classifier inherently learns the structure of each class without further context.

Results for PERSUADE Table 3 shows the performance of both token and sentence classifiers for PERSUADE. Similar to DARIUS, we observe substantial performance variations across the benchmark datasets. On complete essays, the token classifier achieves an overall F1 score of .57, with stronger performance on *lead* (.78) and *concluding statement* (.77), probably due to their fixed positions.

For partial texts, the **First_X** benchmarks show a moderate decline in the overall performance (.52-.46). *Lead* detection decreases (.67 to .57), while *evidence* detection improves (.44 to .78), suggesting that positional cues become less reliable, while content-based identification improves with more context. *Concluding statement* detection remains near zero except for the **First_75**, with a score of .14. This is, similar to DARIUS, due to the lack of concluding statements in the first part of the texts. The **Last_X** benchmarks reveal significant performance degradation. Overall, the F1 scores (.11-.25) are much lower than for the **Full text** or **First_X**. Similar to the findings for DARIUS, the *conclusion* shows a sharp decrease (.77 to .39), suggesting that the positional context is also critical for PERSUADE. The weakest performance appears in the lead (.21) and claim (.00) detection in the **Last_25**, where claims are absent or near-absent.

In the decontextualized **Sentence** and **Shuffled** conditions, the token classifier performs poorly (.20 and .30 overall), while the sentence classifier achieves better results (.43 overall), particularly for *position* (.51) and *evidence* (.48).

These results confirm our DARIUS findings: AM models trained on complete essays have difficulty with partial or nonsequential inputs. The more complex argumentation schema in PERSUADE (seven classes versus three in DARIUS) appears to enhance this issue.

5.3 Experiment 2: Training on Incomplete Texts

Our previous experiment demonstrated that AM models trained on complete texts showed a decrease in performance when confronted with incomplete or out-of-context texts in the case of DARIUS, especially for the *conclusion*. To further investigate this issue, we explored whether training on deliberately split texts could enhance the models' robustness on the benchmark datasets.

Experimental Setup We developed two additional training strategies for the DARIUS dataset. Both used the same amount of training data but differently divided:

The **Split** model was trained on randomly split versions of the training texts, similar to the **First_X/Last_X** benchmark. Each essay was divided into complementary portions (25%/75%, 50%/50%, or 75%/25%).

The **Hybrid** model was trained on a combined dataset consisting of both complete essays (as in our original token classifier) and their split versions (as in the **Split** model).

Both models used the same BERT architecture as our original token classifier and only differed in the training data composition. We then evaluated these models on the same benchmark test sets as those used in our previous experiments.

Results and Analysis Table 4 presents the performance differences between our new models (**Split** and **Hybrid**) and the original token classifier across all test conditions. The values represent changes in F1 scores relative to the original model. Several key patterns emerge from these results:

1. **Performance on complete essays**: Both the **Split** and the **Hybrid** models showed slight performance decreases (F1 scores of -.04 and -.02 overall, respectively) when tested on complete essays, with the most substantial drop observed for conclusion detection (-.17 and -.10). This suggests that, when testing on complete essays, models benefit from training on well-formed, complete training essays, as these contain valuable signals for the task.
2. **First_X**: For essays containing only beginning portions, both models performed similarly to the original classifier, with minor decreases in performance (F1 scores of -.01 to -.03 overall). This indicates that detecting introductions and main content remains relatively robust across training strategies.
3. **Last_X**: The most substantial improvements appeared in the **Last_25** condition, where the **Split** model achieved a +.12 increase in the overall F1 score, with an enormous +.40 improvement in conclusion detection. The **Hybrid** model showed more moderate but still positive gains (+.06 overall, +.22 for conclusions). This pattern of improvement continued in the **Last_50** condition but diminished in the **Last_75** as texts become more complete.
4. **Decontextualized conditions**: For completely decontextualized sentences, both new models substantially improved conclusion detection performance (+.49 for **Split**, +.39 for **Hybrid**), while maintaining or slightly improving overall performance. For shuffled texts, the improvements were small but still positive for conclusion detection.

5.4 Discussion

Our findings indicate trade-offs in training approaches for AM in educational settings. The original token classifier performs well on complete essays but has limitations when identifying argumentative elements, particularly the *conclusion*, when these appear in unexpected positions or without sufficient context.

	Token Classifier on Token Testset				Token Classifier on Sentence Testset			
Variant	**F1 overall**	**F1 Intro**	**F1 Main**	**F1 Conc**	**F1 overall**	**F1 Intro**	**F1 Main**	**F1 Conc**
Full text	.92	.88	.95	**.71**	.93	.91	.96	**.73**
Sentence	.82	.70	.93	**.00**	.82	.71	.93	**.00**
Shuffled	.75	.31	.88	.18	.74	.31	**.87**	.17
First_25	.92	**.90**	.94	**.00**	.93	**.92**	.94	**.00**
First_50	**.95**	**.90**	**.97**	.04	**.96**	**.92**	**.97**	.09
First_75	**.95**	.88	**.97**	.54	**.96**	.91	**.97**	.56
Last_25	**.70**	**.05**	**.86**	.30	**.72**	**.07**	**.87**	.32
Last_50	.86	.25	.93	.57	.88	.30	.94	.60
Last_75	.92	.70	.95	.68	.92	.76	.96	.70
					Sentence Classifier on Sentence Testset			
Sentence					.91	.86	.95	.60

Table 2: DARIUS F1 score. Highest and lowest score per column are bold.

	Token Classifier on Token Testset							
Variant	**Overall**	**Lead**	**Position**	**Claim**	**Counterclaim**	**Rebuttal**	**Evidence**	**Conclusion**
Full text	**.57**	**.78**	.58	**.43**	**.48**	**.42**	.64	**.77**
Sentence	.20	.41	**.19**	.06	**.09**	**.00**	.23	.12
Shuffled	.30	.46	.29	.20	.37	.12	**.13**	.22
First_25	.52	.67	**.61**	.39	.39	.16	.44	**.00**
First_50	.51	.60	**.61**	.38	.42	.35	.76	**.00**
First_75	.46	.57	.56	.31	.34	.32	**.78**	.14
Last_25	**.11**	**.21**	.39	**.00**	.50	.21	.41	.39
Last_50	.12	.33	.47	.01	.56	.36	.61	.48
Last_75	.25	.35	.46	.05	.60	.36	.71	.49
	Token Classifier on Sentence Testset							
Variant	**Overall**	**Lead**	**Position**	**Claim**	**Counterclaim**	**Rebuttal**	**Evidence**	**Conclusion**
	Sentence Classifier on Sentence Testset							
Sentence	.43	.39	.51	.39	.47	.26	.48	.30

Table 3: PERSUADE F1 score. Highest and lowest score per column are bold.

	Overall F1			**Conclusion F1**		
Variant	**Orig.**	**Split**	**Hybrid**	**Orig.**	**Split**	**Hybrid**
Full text	.92	.88 (-.04)	.90 (-.02)	.71	.54 (-.17)	.61 (-.10)
First_25	.92	.90 (-.02)	.91 (-.01)	.00	.00 (.00)	.00 (.00)
First_50	.95	.93 (-.02)	.94 (-.01)	.04	.00 (-.04)	.02 (-.02)
First_75	.95	.93 (-.02)	.94 (-.01)	.54	.40 (-.14)	.43 (-.11)
Last_25	.70	**.82 (+.12)**	**.76 (+.06)**	.30	**.70 (+.40)**	**.52 (+.22)**
Last_50	.86	.87 (+.01)	.87 (+.01)	.57	**.68 (+.11)**	**.62 (+.05)**
Last_75	.92	.89 (-.03)	.90 (-.02)	.68	.62 (-.06)	.63 (-.05)
Sentence	.82	**.86 (+.04)**	**.86 (+.04)**	.00	**.49 (+.49)**	**.39 (+.39)**
Shuffled	.75	.71 (-.04)	.74 (-.01)	.18	**.22 (+.04)**	**.19 (+.01)**

Table 4: Comparison of training strategies on DARIUS dataset. Parentheses show differences from the original classifier. Positive values indicate improvement. Bold values are substantial improvements.

The **Split** model shows that training on incomplete texts improves the handling of such conditions, although this affects the performance on complete essays. The **Hybrid** model offers a trade-off, making modest improvements for incomplete texts while largely maintaining the performance level on complete essays.

The most notable improvements in both new models relate to conclusion detection in partial texts; this is consistent with our observation that conclusions tend to be context-dependent. Training models with conclusions in various contexts reduces their reliance on position and directs attention more to the linguistic features of conclusive statements.

6 Conclusions and Future Work

The results suggest that providing feedback during the writing process is feasible, particularly when students write linearly from beginning to end, but struggles with incomplete or out-of-sequence texts. Context plays a crucial role in accurate classification. While sentence-level classification shows promise for handling incomplete texts, conclusion detection remains challenging as it appears most context-dependent. Our **Hybrid** training approach offers a practical compromise, showing modest improvements for incomplete texts while largely maintaining performance on complete essays. This suggests that educational feedback systems could provide reasonably accurate feedback throughout the writing process while maintaining acceptable performance on complete essays. However, a challenge remains in determining which model to use in real time. As we cannot know a priori whether a student will submit a complete or partial text or whether they have written it sequentially, selecting the optimal model becomes difficult.

Future work could explore methods for detecting completion stages of student texts, enabling dynamic model selection. Additionally, we plan to use process data such as key logs to better understand the writing process and when to provide appropriate feedback. Large language models could be used to create realistic examples of incomplete student essays for training, opposed to only truncated texts.

Limitations

We presented a preliminary study that aims to emulate learner texts in the progress of being written. We were not able to verify how learners actually write in key-logging data; this will be one of our next steps. Thus, our experiments assess possible worst-case scenarios of what incomplete texts might look like, where exactly learners are on this spectrum, is yet to be determined.

Our experiments focus on BERT-based models, chosen for their established performance in German educational contexts and bidirectional attention capabilities. Another option could be to use decoder-based models like DeBERTa. However, such models only look at the left context and thus might not be optimal in scenarios like our where potentially the first part of an essay is still missing

Another limitation of our study is its restrictions to German and English data from Germany and the United States, limiting our finding to two well-resourced languages and only two education systems. More research targeting other languages and datasets would increase the transferability of our results.

Ethics Statement

Our datasets do not contain any new material for which we have to ensure data protection and the handling of personally identifiable information. We selected datasets that, to the best of our knowledge, handled such issues with care.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-document Transformer. *arXiv preprint arXiv:2004.05150*.

S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

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

Yuning Ding, Marie Bexte, and Andrea Horbach. 2022. Don't drop the topic - the role of the prompt in argument identification in student writing. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 124–133, Seattle, Washington. Association for Computational Linguistics.

Yuning Ding, Julian Lohmann, Nils-Jonathan Schaller, Thorben Jansen, and Andrea Horbach. 2024. Transfer learning of argument mining in student essays. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 439–449, Mexico City, Mexico. Association for Computational Linguistics.

Yuning Ding, Ruth Trüb, Johanna Fleckenstein, Stefan Keller, and Andrea Horbach. 2023. Sequence tagging in EFL email texts as feedback for language learners. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 53–62, Tórshavn, Faroe Islands. LiU Electronic Press.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271, New Orleans, Louisiana. Association for Computational Linguistics.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024a. Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 210–221, Mexico City, Mexico. Association for Computational Linguistics.

Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024b. DARIUS: A comprehensive learner corpus for argument mining in German-language essays. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4356–4367, Torino, Italia. ELRA and ICCL.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Maja Stahl, Nadine Michel, Sebastian Kilsbach, Julian Schmidtke, Sara Rezat, and Henning Wachsmuth. 2024. A school student essay corpus for analyzing interactions of argumentative structure and quality. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2661–2674, Mexico City, Mexico. Association for Computational Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. Al: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Thiemo Wambsganß, Tobias Kueng, Matthias Söllner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA. Association for Computing Machinery.

A Appendix

Variant	#Token	#Intro	#Main	#Conc.
Full text	155,502	18,346	123,477	13,679
Sentence	155,502	18,346	123,477	13,679
Shuffled	155,502	18,346	123,477	13,679
First_25	45,600	16,258	29,342	0
First_50	81,753	17,792	63,931	30
First_75	122,472	18,164	102,088	2,220
Last_25	47,203	351	34,203	12,649
Last_50	83,452	834	68,939	13,679
Last_75	124,225	4,082	106,464	13,679

Variant	#Sent.	#Intro	#Main	#Conc.
Full text	8,296	1,150	6,464	682
Sentence	8,296	1,150	6,464	682
Shuffled	8,296	1,150	6,464	682
First_25	2,411	1,013	1,398	0
First_50	4,377	1,121	3,254	2
First_75	6,575	1,142	5,333	100
Last_25	2,411	16	1,758	637
Last_50	4,377	48	3,647	682
Last_75	6,575	264	5,629	682

Table A.5: Count of tokens and sentences for Introduction, Main Part and Conclusion in DARIUS Benchmarks.

- *lead:* Opening hook that guides to thesis
- *position:* Core argument on the topic
- *claim:* Supporting point for position
- *counterclaim:* Opposing viewpoint
- *rebuttal:* Defense against counterclaim
- *evidence:* Support for any argument
- *concluding statement:* Summarizing paragraph

Table A.6: The argumentative components of Crossley et al. (2022)

Variant	Example
full text	CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years. But how can we do this work as efficiently as possible to meet the energy demand in this district? There are three options to choose from. The construction of a wind energy park , a solar plant or a hydroelectric power plant . Which of these three projects should be supported ? if one considers the first criterion, i.e. efficiency, the hydropower plant is clearly in the lead with 70 - 90% . The solar park is the weakest here . The second criterion, annual yield, is close because both the wind farm and the hydropower plant have good annual yields . And so it goes on with the other criteria so that one can say that the construction of a hydropower plant is the most efficient in relation to the construction of an energy converter . The only major drawback is the price of the project, which brings us back to the wind farm because the price difference alone is a whopping 55 million euros . And the criteria are not so bad that you can't compensate for the deficits over the years. In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large .
sentence	CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years.
shuffled	Which of these three projects should be supported ? There are three options to choose from. The construction of a wind energy park , a solar plant or a hydroelectric power plant . And the criteria are not so bad that you can't compensate for the deficits over the years. But how can we do this work as efficiently as possible to meet the energy demand in this district? In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large . The second criterion, annual yield, is close because both the wind farm and the hydropower plant have good annual yields . And so it goes on with the other criteria so that one can say that the construction of a hydropower plant is the most efficient in relation to the construction of an energy converter . The only major drawback is the price of the project, which brings us back to the wind farm because the price difference alone is a whopping 55 million euros . If one considers the first criterion, i.e. efficiency, the hydropower plant is clearly in the lead with 70 - 90%. The solar park is the weakest here . CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years.
first_25	CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years. But how can we do this work as efficiently as possible to meet the energy demand in this district?
first_50	CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years. But how can we do this work as efficiently as possible to meet the energy demand in this district? There are three options to choose from. The construction of a wind energy park , a solar plant or a hydroelectric power plant .
first_75	CO2 emissions and greenhouse gas emissions are to be drastically reduced as they have had a major impact on our climate change in recent years. But how can we do this work as efficiently as possible to meet the energy demand in this district? There are three options to choose from. The construction of a wind energy park , a solar plant or a hydroelectric power plant . Which of these three projects should be supported ? if one considers the first criterion, i.e. efficiency, the hydropower plant is clearly in the lead with 70 - 90% . The solar park is the weakest here .
last_25	And the criteria are not so bad that you can't compensate for the deficits over the years. In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large .
last_50	The second criterion, annual yield, is close because both the wind farm and the hydropower plant have good annual yields . And so it goes on with the other criteria so that one can say that the construction of a hydropower plant is the most efficient in relation to the construction of an energy converter . The only major drawback is the price of the project, which brings us back to the wind farm because the price difference alone is a whopping 55 million euros . And the criteria are not so bad that you can't compensate for the deficits over the years. In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large
last_75	Which of these three projects should be supported ? if one considers the first criterion, i.e. efficiency, the hydropower plant is clearly in the lead with 70 - 90% . The solar park is the weakest here . The second criterion, annual yield, is close because both the wind farm and the hydropower plant have good annual yields . And so it goes on with the other criteria so that one can say that the construction of a hydropower plant is the most efficient in relation to the construction of an energy converter . The only major drawback is the price of the project, which brings us back to the wind farm because the price difference alone is a whopping 55 million euros . And the criteria are not so bad that you can't compensate for the deficits over the years. In conclusion, it can be said that due to the enormous price, a wind farm is the best way to produce electricity for a region that is not so large .

Table A.7: Examples for each category in our benchmark dataset based on essay 943_n3 in the DARIUS dataset (English translation provided by the authors)