Predicting Functional Content Zones in German Source-Dependent Argumentative Essays: Experiments on a Novel Dataset

Xiaoyu Bai and Manfred Stede
Applied Computational Linguistics
University of Potsdam
Karl-Liebknecht-Straße 24-25
14476, Potsdam, Germany
{xiaoyu.bai|stede}@uni-potsdam.de

Abstract

We present a crowdsourced corpus of German-language essays that addresses a middle-school level argumentative writing task. The essays have been composed in response to news articles as reading material, with instructions for the writer to make reference to arguments in the source text in a structured way. The essays were then annotated with fine-grained sentence-level content zones that reflect sentences' functional roles in the essay. Our long-term goal is training a tutoring tool to automatically predict such content zones in unseen essays and to generate helpful feedback on them. To this end we experimented with a range of machine learning models for the prediction task, including a large language model. We obtained the best performance from an ensemble model.

1 Introduction

A common argumentative essay writing exercise in German secondary school is known as *textgebundene Erörterung*[1] (TE), roughly "text-bound argumentation". Students are asked to read an article that discusses a social topic and subsequently to write an essay in which they introduce the topic, summarise and analyse the article author's argumentation and conclude with thoughts of their own. In the long term, we aim to create an intelligent tutoring system (ITS) that supports students in this exercise by providing formative feedback on their essay drafts. Specifically, the system will check whether or not an essay contains the core components of a typical TE-like essay, such as an introduction, paraphrases of arguments from the article, novel arguments of one's own etc. and will point out their absence to the student writer and recommend a revision in a targeted way.

Der Zeitungsartikel "Twitter-Unterricht!" von Tomasz Kurianowicz erschien am 22.06.2011 in der "Zeit Online" (online). Darin behandelt der Autor die heutzutage als seiner Sicht positive Wirkungen von der Benutzung von Soziale Medien als Unterrichtshilfsmittel.
Der Autor weist auf die Benutzung von Twitter als Diskussionforum. Schülern können sich leichter an dem Unterricht beteiligen, indem sie 140 Zeichen Kommentare zu einem Unterrichtsthema aufladen.[...]
Laut dem Autor gibt es zurzeit nur ein sehr kleiner Anteil von Lehrern in den Verinigten Staaten die Twitter als Unterrichtsmedium verwenden und mindestens 50% die skeptisch auf das Internet sind.[...]
Auf der anderen Seite erläutert der Autor dass, der positive Effekt von Twitter für die Gesprächkultur wurde von Psychologen bewiesen sei und obwohl man sich von der sozialen Netzwerken früher leicht ablenken konnte, haben sich die Möglichkeiten mittlerweile verändert.[...]
Im Großen und Ganzen kann ich die Argumentation des Autors gut nachvollziehen und meiner Meinung nach können die Sozialen Medien die Unterrichtsqualität verbessern indem sie interaktiver das Unterrichtsthema darstellen aber durch die verschiedenen Pop-up Werbungen, die man mittlerweile auf allen großen Seiten finden kann, könnten die Schüler abgelenkt werden.[...]

Content zones
info_intro, article_pro, article_con, own

Figure 1: Partly shortened example essay from our corpus with content zone annotations. Labels used in the annotation are explained in the text.

For this purpose, our present work presents a novel, crowdsourced dataset consisting of TE-like essays that have been written on three discussion topics, based on three news articles as source text. To our knowledge, it is the first openly available dataset of German-language argumentative essays that are source-dependent, i.e. written in response to and with reference to a source text. Following existing work which have annotated scientific papers (Teufel and Moens, 2002), parliamentary speeches (Peldszus and Stede, 2016) and student essays (Schaller et al., 2024) with rhetorical or content zones, we annotated our essays with a fine-grained set of content zones, where each zone reflects a functional component of TE-like essays. To illustrate, Figure 1 shows an authentic, albeit partly shortened example essay from our corpus. The essay on the topic of whether or not social media should be used as a tool in schools has been annotated with the following content zones which

[1] See for instance the following prominent tutoring websites: https://www.schuelerhilfe.de/online-lernen/2-deutsch/976-textgebundene-eroerterung and https://www.studienkreis.de/deutsch/textgebundene-eroerterung-schreiben/

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 66–79, 2025

denote the respective functional components:

- **info_intro**: Introductory statement presenting the topic and/or providing (publication) information on the source article
- **article_pro**: Arguments from the source text in support of a given position
- **article_con**: Arguments from the source text against a given position
- **own**: Writer's own argument or viewpoint

For the full dataset, we defined a total of eight content zone labels, which are described in detail in Section 3. Based on our annotated data, we trained machine learning models to automatically predict content zones in unseen essays. Given the source-dependent nature of our task, recognising similarities between arguments in the essays and those in the source text is a particular challenge in our task that sets it apart from the analysis of stand-alone essays.

Overall, the contributions of our present work are the following:

- Through crowdsourcing, we collected a set of 117 German source-dependent argumentative essays written in reference to three news articles.
- We annotated these essays with a pre-defined set of functional content zones.
- We present a series of machine learning experiments for the automatic prediction of these content zones, including using large language models (LLMs). We show that ensembling different classifiers is most successful by a sizeable margin, whereas (smaller) LLMs do not excel.

We make our annotated data and code openly available.[2] Our dataset is referred to as *GerTE* (**Ger**man **T**extgebundene **E**rörterung).

2 Related Work

2.1 Content Zone Recognition

Earlier work on content zone recognition has largely been on domains other than student essays: Working on German parliamentary speeches that argue either in favour of or against a political action, Peldszus and Stede (2016) assign each discourse segment in a speech to one of a pre-defined set of functional building blocks (*content zones*) for the text genre, such as "central thesis" or "pro argument". An older but related approach is taken by Teufel and Moens (2002), who work on scientific papers and annotate them with so-called *rhetorical zones*. They assign each sentence in a given scientific paper to a set of rhetorical classes that are typically found in academic papers, such as research aims or methodology descriptions.

The German-language DARIUS corpus (Schaller et al., 2024) is the only work we know of to apply content zone annotations to student essays. The corpus comprises over 4500 argumentative essays by secondary school students and is annotated with multiple layers of information, most of which target argumentation mining. Content zone annotations for DARIUS use three coarse-grained labels: *Introduction*, *Main Part* and *Conclusion*. Unlike DARIUS, our essays are source-dependent; our content zone labels, which are more fine-grained, target making adequate reference to the argumentative source text rather than independent argumentation.

2.2 Source-Dependent Student Essays

Among the most prominent work addressing source-dependent essays is eRevise (Zhang et al., 2019; Zhang and Litman, 2020, 2021), a tutoring system that assesses secondary student essays' explicit reference to relevant passages from the source text and gives feedback on them. Other than that, the highly popular Kaggle ASAP-AES dataset[3], which has been used in a vast amount of automated essay scoring research (Bai and Stede, 2023), also features source-dependent essays (Mathias and Bhattacharyya, 2018). However, they are labelled for holistic scoring, which does not specifically target reference to the source text.

For German, accessible datasets of school-level essays are scarce. In addition to DARIUS (Schaller et al., 2024), Stahl et al. (2024) have released a dataset of secondary-level students' persuasive essays with argumentation annotations. However, neither corpus contains source-based essays. Horbach et al. (2017) have scored source-based, German-language summary essays, yet the dataset is drawn from university students and is not publicly available. We are not aware of work on

[2] https://github.com/XYB1001/GerTE-Konvens-25

[3] https://www.kaggle.com/competitions/asap-aes

German that addresses source-dependent secondary school-level essays similar to eRevise (Zhang et al., 2019) for English. We address this gap by releasing our present dataset.

3 Data

3.1 Data Collection

Data collection from real school students is difficult due to legal and ethical concerns. Therefore, as a second-best solution for starting the new task, we used crowdsourcing to collect essays in response to our TE exercise. We chose three news articles that deal with the following easily accessible discussion topics, each with a short-hand name given in brackets:

1. Should Twitter be integrated into school classes as a tool?[4] (*Twitter*)
2. Should climate change be taught at school in a subject of its own?[5] (*Climate*)
3. Should school start later in the morning?[6] (*LateSchool*)

The articles are openly accessible and approximately 600 - 700 words in length. Along with a link to the article and an example essay, an instruction with approximately the following wording (albeit in German) is given to the crowd workers: *Read this news article. Write a "textgebundene Erörterung" (TE) of approximately 300 to 350 words, using the example essay as an orientation if needed. Summarise the topic and the arguments given by the author in a structured way and conclude with your own viewpoint on the topic*. The example essay deals with a fourth topic that is not one of the three topics listed above and was written by the first author of the present paper.

The data collection process took place in April 2021 on the Prolific[7] crowdsourcing platform. We hired workers above the age of 18 who reported a good level of general education and sufficient proficiency in German. Overall, we obtained 117 essays, 50 each for the topics of *Twitter* and *LateSchool* and 17 for the topic of *Climate*, which was considered a less straightforward topic. All submissions were rigorously checked for quality, and those of low quality were replaced with new submissions. The essays range between 210 to 445 words, with an average length of 267.6 words. Statistics divided by topic are shown in Table 1. For illustration, one of the essays for the topic of *Twitter* is shown in full in Figure 5 in the Appendix.

Topic	Twitter	LateSchool	Climate	Total
Range	210 - 313	215 - 445	230 - 395	210 - 445
Mean	249.5	268.8	317.7	267.6
SD	25.5	39.8	40.8	41.0

Table 1: Word count statistics for the essays on each topic in our collected dataset.

3.2 Data Annotation

In our approach to recognising core components of a TE-like essay, we follow Teufel and Moens (2002) and work on the sentence level, assigning each sentence in an essay to a single functional content zone (cz). We define a total of eight fine-grained cz-labels, which are summarised in Table 2. The example sentences shown in the table are based on authentic data from our dataset (translated and in part shortened).

With regard to the *LateSchool* topic, where no clear stance is put forward by the article's author, the label *article_pro* is used for sentences which support starting school later and *article_con* for sentences arguing against it. We consider every token to be part of a sentence (whether well-formed or not) and therefore as belonging to one of these cz classes. Hence, the label *other* is introduced as a fall-back class for any chunk of text that cannot be assigned to any of the other seven classes, including badly-formed and/or incomprehensible pieces of text.

Spacy (Honnibal and Montani, 2017) was used to perform sentence segmentation. Two annotators – the first author of the present work and a research assistant – each independently annotated all 117 essays and achieved a good inter-annotator agreement of 0.78 (Cohen's Kappa). In an adjudication process, all annotations showing disagreement were discussed and a single annotation agreed on.

Out of the 1682 essay sentences, 64 (3.8 %) were temporarily assigned to two different cz-classes as

[4]From Zeit-Online, at https://www.zeit.de/digital/internet/2011-06/twitter-unterricht/komplettansicht, last access on 30.04.2025

[5]From Zeit-Online, at https://www.zeit.de/gesellschaft/schule/2020-01/klimawandel-schulfach-bildung-unterricht-konkurrenz, last access on 30.04.2025

[6]From ÄrzteZeitung Online, at https://www.aerztezeitung.de/Panorama/Ist-es-vernuenftig-die-Schule-um-8-zu-beginnen-402238.html, last access on 30.04.2025

[7]https://www.prolific.co/

Cz-label	Description	Example
info_article	Sentences presenting information on the source article, such as author, publication date etc.	The news article "Twitter-Unterricht" was published on 22nd of June, 2011 in "Zeit" (online).
topic_intro	Sentences which introduce the discussion topic	In the article, the author discusses whether it makes sense that school starts as early as 8:00 in the morning.
article_pro	Sentences from the article that *support* the author's side of the argumentation	The author believes that Twitter allows students to better get in touch with their teachers.
article_con	Sentences from the article that *undermine* the author's viewpoint	The authors sees the only counter-argument in old fashioned teachers who are sceptical towards new media.
own	Sentences which present the student writer's own position or thoughts on the discussion topic	I believe the author's suggestion is a good idea and would really help the educational system.
meta	Sentences that the writer uses to structure the essay	Now I turn to the arguments against this idea.
off_topic	Sentences that are comprehensible but do not relate to the discussion topic and feel "off topic"	Recently young people have forgotten the past, which has given rise to issues like racism [found in an essay on the Climate topic] .
other	Fall-back class for sentences for which none of the above content zone labels apply	Also the news from the world, and it's the last minute [found in an essay on the Twitter topic].

Table 2: Content zones used in our annotation scheme

neither can be said to be more significant:

- 33 sentences were labelled with both *info_article* and *topic_intro* and are of the form *The article, which was published in ... deals with the question of ....* Since both content zones form introductory statements for our essays and often can be expressed by a single sentence, we chose to conflate the two classes into a single class *info_intro*.

- The remaining cases of double labelling are of the form *Although [pro/con argument from article], [pro/con argument from article or own argument].* An authentic (but translated) example from our data reads *Although children would like to sleep longer and go to school later, they also like it when school finishes earlier, which gives them much free time in the afternoon.*. We manually split these sentences into sentential clauses such that each clause could be annotated with a single cz-label. They were thus treated as separate single-label sentences.

After these steps, our overall dataset for content zone recognition consists of 1713 sentences from 117 essays, each sentence being labelled with exactly one out of seven possible cz-labels. The distribution among the seven classes is shown in Figure 6 in the Appendix.

4 Classification Experiments

In our dataset, samples for the classes *meta*, *off-topic* and *other* are significantly less frequent than the remaining classes, with less than 100 samples for each of them in the full dataset. Moreover, sentences of these classes are less relevant to our prospective use case where we wish to draw attention to missing core components of an TE-like essay, which these classes do not count among. Due to these reasons, in our current classification experiments we conflated the classes *meta*, *off-topic* and *other* into a single class *other*. The resulting cz label distribution based on five classes in total is as shown in Figure 2. This is the data basis for our classification experiments.

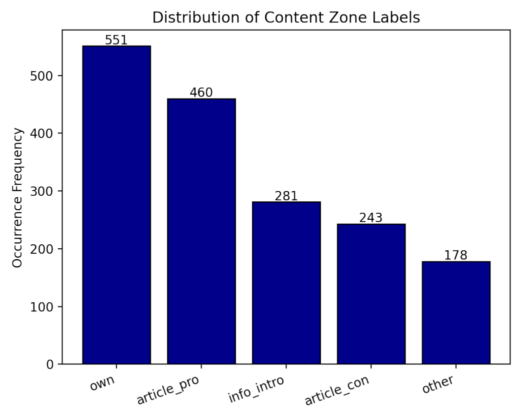

Figure 2: Distribution of content zone labels in our data in the conflated 5-class setting.

With the exception of the experiments based on

LLMs, all of our experiments were carried out in 5-fold cross-validation. In each fold, the models were trained on roughly 80% of the essays (93 or 94 out of 117) and tested on the held-out 20%.

4.1 Preprocessing

Apart from sentence segmentation, which had already been done prior to annotation, we did not undertake any preprocessing steps.

4.2 Basic Feature-Based Models

We experimented with a combination of feature-based and neural, representation-based models. All feature-based machine learning models used implementations by scikit-learn (Pedregosa et al., 2011). Our baseline was either a linear SVM or a random forest classifier using only TF-IDF-weighted word unigrams and bigrams. Additional hand-crafted features we experimented with are the following:

Surface character features A straight-forward addition to the baseline is the addition of character n-grams in the range [3,4].

Position features We defined position features to encode the relative position of a given sentence in the essay that it was drawn from. For this, we split each essay into four parts, where all parts have approximately the same number of sentences. The position feature is defined as 4 binary features with values 1 or 0, where each represents one of the four quarters of the essay and 1 indicates that a given sentence occurs in the corresponding quarter. This feature aims at capturing the intuition that some content zones (e.g. introduction statements) tend to occur at specific positions within the essay. For instance, in our dataset, 252 out of the 281 instances (89.7%) of *info_intro*-labelled sentences occur within the first three sentences of the essay; 104 out of the 117 essays (88.9%) end in a sentence labelled as *own* as most writers concluded their essays with a final verdict of their own on the discussion topic.

Static BERT embedding features We experimented with static embeddings from pre-trained BERT (Devlin et al., 2019), specifically the German model "bert-base-german-cased"[8], released by deepset.ai[9]. For each sentence, we took the representation of the special `[CLS]` token as its embedding.

SBERT article similarity features We used sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to encode the amount of similarity between a target sentence and each sentence in the corresponding source article. The motivation here is that sentences labelled as *article_pro* or *article_con* reproduce arguments from the source article and therefore can be expected to be near-paraphrases of and show high similarity scores with specific sentences in the article that introduce those arguments. For instance, for *LateSchool*, a translated essay sentence labelled *article_pro* is: *They have more time for sleep*, which we expect to exhibit a higher similarity score with the article sentence *On average, students slept 34 minutes longer* than with other sentences in the article. SBERT is known to perform well in paraphrase detection tasks (Khairova et al., 2022; Peng et al., 2022) and therefore chosen for this feature. We used the pre-trained multilingual model "sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2"[10], and the cosine measure to encode the similarity between target sentences and each corresponding article sentence.

4.3 Fine-tuning a BERT Model

Alongside feature-based models, we also tested the same BERT-based neural model in the fine-tuning scenario. That is, we again fed each target sentence to BERT and extracted the embedding of the `[CLS]` token. To inject information on the relative position of a sentence within an essay, we concatenated this sentence embedding with our position features as detailed above. The resulting representation was then passed to an output layer with logged softmax for classification.

The model was implemented with PyTorch (Paszke et al., 2019) and Hugging Face (Wolf et al., 2020). We used the PyTorch implementation of the weighted Adam optimiser (Loshchilov and Hutter, 2017) with default parameters and the negative log likelihood loss as our loss function. Based on a preliminary examination using 10 essays (4 from *Twitter* and *LateSchool* and 2 from *Climate*) as validation data, we decided to fine-tune the model for 4 epochs, with a batch size of 16 and a learning

[8] `https://huggingface.co/bert-base-german-cased`

[9] `https://www.deepset.ai/german-bert`

[10] Available through the Hugging Face Hub at `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`

rate of $1e^{-5}$. The final fine-tuning experiment was again carried out in 5-fold cross-validation.

4.4 Ensemble Modelling

As an alternative to using n-grams and embeddings as different feature sets in order to exploit both surface-level and deep representational information, we experimented with combining them through ensemble modelling. In this approach, multiple base models learned to output probability values for each cz class, and a meta model predicted the final cz label based on these probability features. We chose the following three base models, which used different features that have already been described above:

- Model 1: linear SVM with static BERT embeddings, SBERT article similarity and position features
- Model 2: linear SVM with word and character ngrams, SBERT article similarity and position features
- Model 3: Random Forest with word and character ngrams, SBERT article similarity and position features

At training time, we used 10-fold cross-validation within the training partition of the overall 5-fold cross-validation scheme to obtain content zone predictions by the three models on the training partition[11]. Each prediction is a probability distribution among the 5 possible content zone labels. This gave us a set of base model predictions on each sentence in the training partition.

At the meta level, we used a logistic regression model which took the base predictions for each sentence as its sole feature, i.e. a vector of 15 components (5 probability values from each of the 3 base models), and finally produced a label for the sentence. At training time, the meta-model trained on the base model predictions for the sentences in the full training partition, paired with the true content zone labels in the annotations. At test time, we retrained each of the three base models on the full training partition and applied them to the test data. Their probability predictions for the test sentences were then fed to the trained meta-model to obtain the final labels for the test partition.

[11]That is, rotating through 10 folds, they repeatedly trained on 90% of the training partition and made predictions for the remaining 10%.

4.5 Prompting an LLM

As a final approach, we experimented with prompting an LLM to output cz labels. We chose Meta's instruction fine-tuned Llama3.1 model with 8 billion parameters[12] (Grattafiori et al., 2024) as it is an openly available, state-of-the-art multilingual LLM that is still easy to use in terms of hardware requirements. Access to the model was through its integration with the Groq[13] framework. We examined three prompting scenarios:

- **Zero-shot**: We treated all of our data as test data and prompted Llama to output content zone labels for each essay in a single run. The prompt did not include the source article text.
- **Zero-shot with article**: Same as *zero-shot*, but we included the article text which the target essay references in the prompt.
- **One-shot**: We manually inspected the labelled essays and chose a single essay[14] to include in the prompt as a demonstration for in-context learning. The chosen essay deals with the *Twitter* topic and well exemplifies each of the five content zone labels. We then used this one-shot prompt to ask Llama to produce labels for all other essays in the dataset; the article text was not included in this setting.

A gloss of the system prompt we used is: *You are a middle-school German teacher who checks whether a student's essay contains all the expected essay components.* The user prompt approximately reads: *Here is an argumentative essay written by a student in response to reading a text. The essay is split into sentences. Each sentence is numbered and follows this format [...]. Each sentence can be paired with one of the following five functions.* Following a concise description of each content zone label using one sentence, the prompt concludes with *Assign each sentence to one of these five functions. The output should follow this format [...]* before displaying the target essay in the format promised. In the *zero-shot with article* scenario, we appended the system prompt with the sentence *Here is the article to*

[12]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[13]https://groq.com/
[14]The essay chosen to serve as demonstration and its gold content zone labelling are given in the Appendix.

which the essay refers and the full article text. In the prompt texts, we gave the content zone labels meaningful German names, such as *Einleitung* or *Pro_aus_Lesetext* for *info_intro* and *article_pro*, respectively. An example of our original German-language prompts is provided in the Appendix.

In its output, Llama was prompted to reference each sentence by its number and pair it with a cz label. This allowed us to extract the labels from the output text using simple Regex-based rules. If the number of labels extracted matched the number of sentences in the essay (and thus the number of gold labels), we could apply the same evaluation procedure as for all the other classification models.

4.6 Sentence-Level Sequence Classification

Apart from the LLM-based approach, all of the models described above cast the task as a sentence classification task[15] where a single sentence is independently labelled with a single content zone. The handcrafted position feature detailed above is the only indication concerning a sentence's relative position to other sentences from the same essay.

As an alternative, we also experimented with casting our task as a sequence labelling task on the sentence level. In this case, each essay is represented as a sequence of sentence representations and labelled with a corresponding sequence of content zones. We experimented with a BiLSTM-CRF model architecture (Huang et al., 2015) with different sentence embeddings. However, our initial experiments showed very poor classification results; we therefore did not further pursue this approach. Since this approach works on the essay instead of the sentence level, we suspect that our dataset of 117 essays (of which less than 100 are used for training) is too small to perform efficient training on.

5 Results

We evaluated our models in terms of the common classification metrics accuracy, precision, recall and F1. For the latter three, we report the weighted macro average across the five classes. The most frequent classes in our dataset, viz. those with the labels *article_pro*, *article_con* and *own*, are also those of the most interest to our intended use case of recognising central components of a TE-like essay. It therefore appears plausible to weigh the classes by their frequency. Table 3 shows the results in terms of these metrics. With the exception of the Llama models, the scores for all models are averaged across the five folds in cross-validation.

Regarding the Llama models, in the two zero-shot scenarios, 12 (without the article in the prompt) and 8 (with the article) essays, respectively, led to output in which the numbers of output labels and sentences did not match and which were therefore excluded from evaluation. In the one-shot scenario, 10 essays produced invalid output. For each prompt scenario, we report the single set of metrics as evaluated on all the essays in the respective scenario that produced valid output and were considered in evaluation. In line with expectations, both the addition of the source article and of an appropriate demonstration essay boosts the LLM's performance (though the improvement from the one-shot scenario is minor).

6 Discussion and Future Directions

Overall, the ensemble model clearly produced the best results. The fine-tuned BERT model and the hybrid model using an SVM with both surface features and static BERT embeddings also achieved competitive results. Moreover, the simple and fast approach of feeding word and character ngrams with position features to an SVM turned out to be remarkably successful, achieving comparable performance with fine-tuned BERT.

To further analyse the results, we show in Figure 3 the by-class F1 metric of the following three models, as averaged across five folds: the baseline SVM, the SVM with surface and static BERT features (*Hybrid*) and the best-performing ensemble model. The class *other* clearly stands out as the worst-performing class across the board and the only class on which the overall best-performing ensemble model does *not* achieve the highest F1 score. The low performance is expected since the class is both heavily under-represented and fuzzily defined since it serves as a fall-back class for sentences that do not fit elsewhere. However, with respect to our designated use case, it is also the least relevant class. The easiest class to recognise is clearly *info_article*. This is also unsurprising; not only are certain phrases such as *the article deals with...* or *the article was published in...* highly indicative of the class, sentences of this class also almost exclusively occur at the beginning of an essay, which is captured by the position feature.

[15]The LLM-based approach treats the task as a chat-completion task.

	Accuracy	P (weighted)	R (weighted)	F1 (weighted)
SVM + word ngrams (baseline)	0.653	0.683	0.653	0.634
SVM + w&c ngrams + position	0.680	0.679	0.680	0.674
RF + w&c ngrams + position	0.671	0.725	0.671	0.641
SVM + w&c ngram + BERT (static) + position + article	0.686	0.685	0.688	0.682
RF + w&c ngram + BERT (static) + position + article	0.646	0.66	0.646	0.61
BERT (fine-tuned) + position	0.686	0.686	0.686	0.674
Ensemble model	**0.728**	**0.723**	**0.728**	**0.713**
Llama (zero-shot)	0.654	0.636	0.654	0.641
Llama (zero-shot with article)	0.673	0.659	0.673	0.661
Llama (one-shot)	0.661	0.639	0.661	0.649

Table 3: Evaluation results for selected models; *P* and *R* refer to precision and recall, *RF* denotes the random forest classifier, *w&c ngrams* refers to word and character ngrams, *position* denotes position features and *article* denotes article similarity features. Values for all but the Llama models are averaged across five folds.

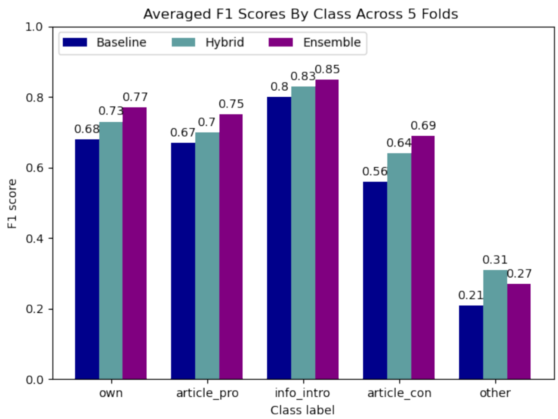

Figure 3: Averaged F1 scores by class for three selected models.

The success of the ensemble model could be attributed to performance behaviour of the different base models that complement each other to some degree, which can be seen in the confusion matrices shown in Figure 4. For instance, Plot b) suggests that the random forest model with n-gram and position features has a bias towards the class *own* such that it correctly recognises the large majority (86%) of them but also overpredicts over 30% of *article_con* and nearly 60% of *other* sentences as *own*. This tendency is not shared by the hybrid SVM with surface and BERT embeddings (Plot a)); in contrast, it predicts (and overpredicts) *article_con* far more often than the random forest model, which often misses this class. The ensemble model (Plot c)) seems to build on the strengths of its base models, both achieving high recall on *own* (comparable to the random forest) while reducing misclassification of *article_con* samples as *own* (comparable to the SVM).

With respect to the performance by Llama, we examined the output texts that we had to discard from evaluation and made the following observation: In many of these cases, Llama generated more pairs of sentence number and content zone labels than there were sentences in the input essay, which seems to indicate model hallucination. To illustrate this, an example output along with the corresponding prompt is given in the Appendix.

It should be noted, however, that our Llama experiments are only a first step. Future work could investigate the performance of the larger versions of Llama and of other LLMs such as DeepSeek-R1 (Guo et al., 2025) or commercial ones like GPT-4 (Achiam et al., 2023). Moreover, the fact that classification has benefitted both from the addition of the source article and from the one-shot prompt scenario is in line with expectations and indicates that in the given task, the LLM can successfully exploit additional information in the prompt to increase its performance. Thus, focus on further prompt engineering efforts will likely yield better classification results. For instance, we plan to look into improved and automatic ways to select demonstration essays for in-context learning in a few-shot prompting scenario.

7 Conclusion

In this work, we have presented a novel German dataset consisting of 117 middle-school level, source-dependent argumentative essays and their annotations with fine-grained, sentence-level content zones. While the essays were collected through crowdsourcing, we examined them closely to ensure their quality. Based on this dataset, we trained different machine learning models, including LLMs, to predict content zones in the essays. Our best-performing model is an ensemble model that combines an SVM and

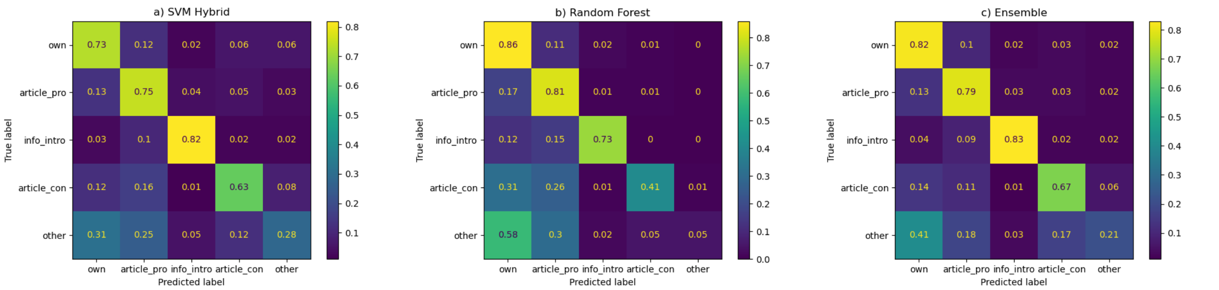

Figure 4: Normalised confusion matrices for three selected models. All values are as averaged across five folds.

a random forest model with both hand-crafted features and static BERT embeddings. We believe further investigation of LLMs can be a promising direction, although we also believe that practical advantages of more lightweight classification approaches remain an important consideration in terms of computation power needed, runtime, energy consumption etc. As the next step, we plan to continue our work by presenting our exercise to real students from schools and collecting authentic data. In the long term, we hope to incorporate our analysis into a tutoring system that generates formative feedback to support students' in their essay writing in TE-like exercises.

Limitations

We point out and acknowledge two limitations with regard to our current dataset: First, with 117 essays it is a fairly small dataset. Second, it has been collected by means of crowdsourcing from adult workers whose essay writing skills are not necessarily representative of middle-school students. We plan to address both limitations in the future by presenting our exercise in schools to extend our dataset with essays from real students.

Ethical Considerations

We do not foresee significant ethical risks in relation to our work. The nature of our exercise and our manual validation of the essays ensure that no personal information is disclosed through the essays. However, since our crowdsourcing data collection was anonymous, we could not ensure that writers of different population groups in terms of gender, ethnicity, socio-economic background etc. were adequately represented. Possible bias in classification models that disadvantage certain population groups should therefore be kept in mind.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiaoyu Bai and Manfred Stede. 2023. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, 33(4):992–1030.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366, Copenhagen, Denmark. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Nina Khairova, Anastasiia Shapovalova, Orken Mamyrbayev, Nataliia Sharonova, and Kuralay

Mukhsina. 2022. Using bert model to identify sentences paraphrase in the news corpus. In *COLINS*, pages 38–48.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Andreas Peldszus and Manfred Stede. 2016. Inhaltszonen. In Manfred Stede, editor, *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0*, pages 133–144. Universitätsverlag Potsdam.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-Argument Based Bi-Encoder for Paraphrase Identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*. ArXiv: 1908.10084.

Nils-Jonathan Schaller, Andrea Horbach, Lars Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024. DARIUS: A Comprehensive Learner Corpus for Argument Mining in German-Language Essays.

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.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Zhang and Diane Litman. 2020. Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8569–8584, Online. Association for Computational Linguistics.

Haoran Zhang and Diane Litman. 2021. Essay Quality Signals as Weak Supervision for Source-based Essay Scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–96, Online. Association for Computational Linguistics.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9619–9625.

A Example Essay Sample

The shortened and annotated essay shown in Figure 1 of the main text is shown in full in Figure 5, without annotations.

B Content Zone Distribution With Seven Classes

The content zone label distribution with seven class labels, before mapping *meta* and *off-topic* to *other*, is shown in Figure 6.

C Example LLM Prompts and Outputs

An example prompt fed to Llama in the zero-shot scenario with an input essay on the *LateSchool* topic is given in Table 4. Table 5 shows the demo essay on the *Twitter* topic and the demo output that we used as our example in the one-shot prompting scenario. Table 6 shows an example of an invalid output on the same topic in the one-shot

Der Zeitungsartikel “Twitter-Unterricht!” von Tomasz Kurianowicz erschien am 22.06.2011 in der “Zeit Online” (online). Darin behandelt der Autor die heutzutage als seiner Sicht positive Wirkungen von der Benutzung von Soziale Medien als Unterrichtshilfsmittel.
Der Autor weist auf die Benutzung von Twitter als Diskussionforum. Schülern können sich leichter an dem Unterricht beteiligen, indem sie 140 Zeichen Kommentare zu einem Unterrichtsthema aufladen. Die niedrige Schwelle, ermöglicht es den Schülern, die normalerweise schweigen würden, am Unterricht Teil zu nehmen.
Laut dem Autor gibt es zurzeit nur ein sehr kleiner Anteil von Lehrern in den Verinigten Staaten die Twitter als Unterrichtsmedium verwenden und mindestens 50% die skeptisch auf das Internet sind. Das passiert wegen der Angst vor Zerstreuung die immer noch sehr groß ist.
Auf der anderen Seite erläutert der Autor dass, der positive Effekt von Twitter für die Gesprächkultur wurde von Psychologen bewiesen sei und obwohl man sich von der sozialen Netzwerken früher leicht ablenken konnte, haben sich die Möglichkeiten mittlerweile verändert.
Der Autor meint dass, die Informationen im Digitalformat besser im Gedächtnis gespeichert werden und die Schüler auch von einem interessanten und abwechslungreichen Unterricht profitieren können und die Einführung des Social Web der Erfindung des Buchdrucks ähnelt und es irgendwann im Unterricht als selbstverständlich betrachten wird.
Im Großen und Ganzen kann ich die Argumentation des Autors gut nachvollziehen und meiner Meinung nach können die Sozialen Medien die Unterrichtsqualität verbessern indem sie interaktiver das Unterrichtsthema darstellen aber durch die verschiedenen Pop-up Werbungen, die man mittlerweile auf allen großen Seiten finden kann, könnten die Schüler abgelenkt werden. Es gibt auch das Problem der immer sinkenden Aufmerksamkeitsspanne die durch Soziale Medien beeinträchtigt wird.

Figure 5: Example essay in full on the *Twitter* topic.

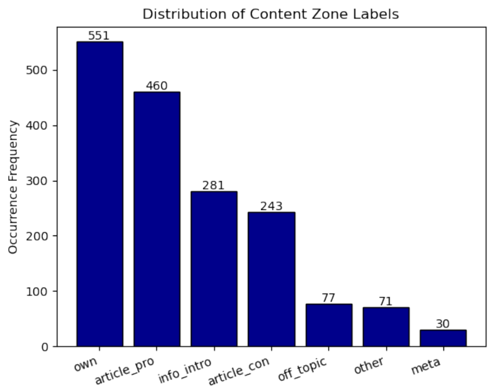

Figure 6: Distribution of our content zone labels in the 7-class setting.

scenario. This output produces 13 pairs of sentence numbers and content zone labels when the essay only consists of 9 sentences. It is therefore not valid and was excluded from evaluation.

Du bist ein Deutschlehrer der 9. Klasse und analysierst Aufsätze deiner Schüler. Du überprüfst dabei, ob die Aufsätze alle zu erwartenden Bestandteile enthält.
Hier liegt ein argumentativer Aufsatz vor, den ein Schüler mit Bezug auf einen Lesetext geschrieben hat. Der Aufsatz wurde in Sätzen unterteilt. Die Sätze sind nummeriert. Das Format dabei is "Satznummer: Satz". Jeder Satz wird einer der folgenden 5 Funktionen zugeordnet: "_Einleitung" bedeutet: Der Satz leitet den Aufsatz ein und gibt evtl. Informationen zum Lesetext und zum Thema des Aufsatzes. "_Pro_aus_Lesetext" bedeutet: Der Satz gibt ein Pro-Argument aus dem Lesetext wider. "_Con_aus_Lesetext" bedeutet: Der Satz gibt ein Kontra-Argument aus dem Lesetext wider. "_Eigen" bedeutet: Der Satz gibt eigene Meinungen und Argumente des Schülers wider. "_Sonstiges" bedeutet: Der Satz kann keiner der vorherigen 4 Funktionen zugeordnet werden. Werte den Schüleraufsatz aus und ordne jeden Satz einer der genannten 5 Funktionen zu. Das Format des Outputs soll ausschließlich wie folgt sein: "Satznummer: Funktion", zum Beispiel "1: _Einleitung" oder "8: _Eigen". Schüleraufsatz 1: In dem Text von Cornelia Dieckmann aus dem Jahr 2019 wird die Frage gestellt, ob der "Schulbeginn um 8 Uhr sinnvoll" ist. 2: Dieser Artikel ist in dem Magazin "Ärzte Zeitung" zu lesen. 3: <PB> Sie greift auf Schlafforschungen zurück, um ein klareres Bild der Vor - und Nachteile zu schaffen. 4: <PB> Laut einer US-amerikanischen Studie werden die Leistungsfähigkeit und die Konzentration der Schüler:innen deutlich verbessert und außerdem schlafen sie ca 30 Minuten länger, wenn der Unterricht um 8:45 beginnnt und nicht um 8:00 Uhr. 5: <PB> Dies bedeutet jedoch auch, dass der Schultag später endet, was Eltern und Schülern nicht gefällt. 6: Denn im Ende geht es um die Freizeit die man am Tag noch hat. 7: Würden die Schulen später anfangen, gäbe es auch mehr Busverbindungs-Probleme, da nun Schulen unterschiedlich beginnen und man trotzdem noch zum pünktlich zum Unterricht muss. 8: <PB> Tageslicht spielt dabei auch eine große Rollle, denn wenn man morgens wenn es noch dunkel ist in die Schule kommt und abend wieder die Schule verlässt, beinträchtigt das zum Ersten die Konzentration und Leistungsfähigkeit der Schüler:innen, und zum Zweiten auch deren Laune. 9: <PB> Zusammenfassend ausgedrückt geht es um den Schüler:innen mehr um den Rest des Tages den sie übrig haben. 10: Denn wenn man eine Stunde mehr draußen,, zum Beispiel am See sein kann, dann nimmt man die frühe Stunde der Schule in Kauf. 11: Die Organisation eines späteren Schulbeginns hört sich am Anfang leicht an, ist aber komplizierter als erwartet was ein Grund dafür ist, die Zeiten beizubehalten. 12: Allerdings darf man die positiven Auswirkungen auf die Schüler:innen nicht igrnorieren, denn nur 30 Minuten mehr schlaf führen zu besseren Ergebnissen.

Table 4: Example prompt fed to Llama in the zero-shot scenario. The top part is the system prompt, the lower part the user prompt.

1: Der Zeitungsartikel "Twitter-Untericht" von Tomasz Kurianowicz erschien am 22.
2: Juni 2011 in der "Zeit Online".
3: Dahin behandler der Autor eine neue Lehrmethode - nutzung der sozial Netzwerken im Unterrichten.
4: Es ist sehr wichtig, dass die Lehrer ihre Studenten zuzugehehen und heutzutage gibt es keine bessere Methode Schülern zu erreichen, als durch sozial Netzwerken.
5: Es ist auch ein guten Trick die Jugendliche in der Unterricht zu angarieren, weil Mehrheit der jungen Leute jeden Tag mehrere sozial Netzwerke nutzt und deswegen ist das spannend für ihnen.
6: Wie alles, hat diese Methode nicht nur Vorteile aber auch Nachteile.
7: Obwohl positive Effekt für die Gesprächskultur durch Psychologen bewiesen sei,
8: viele Lehrer auf diese Method mit skepsis reagieren und nicht verwenden im ihren Unterrichten.
9: Der Grund dafür ist höchstwahrscheinlich die Angst vor Zerstreuung.
10: Tradizionell, können die Schülern während des Unterrichts nicht benutzen, früher Lehrer sogar sammelten die Handys vor dem Unterricht ein, um Ablenkungen zu vermeiden.
11: Für diesen Lehrer ist also twittern im Unterricht fast unvorstellbar.
12: <PB> Persönlich glaube ich, dass heutzutage ist die virtuelle wert gleichermaßen wichtig wie die Realität ist.
13: Manche Leute interagieren mit anderen Persönen durch das Internet sogar mehr, als im Realität.
14: Scheuen sich davon das Internet zu nutzen ist also sinnlos und kann sozial Fähigkeiten der Studenten schaden.
15: Meine Meinung nach, wir sollen neue Lösungen verwenden und nicht nur verlassen sich auf die Tradizion.
1: _Einleitung
2: _Einleitung
3: _Einleitung
4: _Pro_aus_Lesetext
5: _Pro_aus_Lesetext
6: _Sonstiges
7: _Pro_aus_Lesetext
8: _Con_aus_Lesetext
9: _Con_aus_Lesetext
10: _Con_aus_Lesetext
11: _Con_aus_Lesetext
12: _Eigen
13: _Eigen
14: _Eigen
15: _Eigen

Table 5: The example essay and output given to Llama as demonstration in the one-shot scenario. The top part is the input essay, the lower part the demo output.

[Removed for brevity: System prompt, example essay and example output for in-context learning]

Hier liegt ein argumentativer Aufsatz vor, den ein Schüler mit Bezug auf einen Lesetext geschrieben hat. Der Aufsatz wurde in Sätzen unterteilt. Die Sätze sind nummeriert. Das Format dabei is "Satznummer: Satz". Jeder Satz wird einer der folgenden 5 Funktionen zugeordnet:
"_Einleitung" bedeutet: Der Satz leitet den Aufsatz ein und gibt evtl. Informationen zum Lesetext und zum Thema des Aufsatzes.
"_Pro_aus_Lesetext" bedeutet: Der Satz gibt ein Pro-Argument aus dem Lesetext wider.
"_Con_aus_Lesetext" bedeutet: Der Satz gibt ein Kontra-Argument aus dem Lesetext wider.
"_Eigen" bedeutet: Der Satz gibt eigene Meinungen und Argumente des Schülers wider.
"_Sonstiges" bedeutet: Der Satz kann keiner der vorherigen 4 Funktionen zugeordnet werden.
Werte den Aufsatz aus und ordne jeden Satz einer dieser 5 Funktionen zu. Das Format des Outputs soll ausschließlich wie folgt sein: "Satznummer: Funktion", zum Beispiel "1: _Einleitung" oder "8: _Eigen".

Schüleraufsatz

1: Im Artikel von Cordula Dieckmann geht es darum, ob ein späterer Schulbeginn nicht sinnvoller wäre als die aktuelle Form.
2: <PB> Für eine Verschiebung des Beginns spricht, dass den Schülern dadurch mehr Schlafzeit zustehen würde, und somit können sie ausgeschlafen zum Unterricht erscheinen.
3: Da während der Pubertät viele Jugendliche eher dazu neigen, erst später ins Bett zu gehen und aufzustehen, wäre dies eine Möglichkeit, die fehlende Schlafzeit auszugleichen.
4: <PB> Dagegen spricht allerdings, dass die Umsetzung nicht reibungslos erfolgen würde.
5: Zum einen müssen sämtliche Verkehrsdienste wie Schulbusse umstrukturiert werden, um Schulen mit abweichenden Uhrzeiten zu bedienen, zum anderen wäre in so einem Fall zusätzliche Unterrichtsblöcke erforderlich, die den Nachmittag belegen würde.
6: Dies sei unter vielen eine unbeliebte Idee, da sie u.a. im Winter im Dunkel zur Schule gehen und sie auch im Dunkel verlassen würden -
7: Kontakt zum Sonnenlicht regelt nämlich die innere Uhr und ist vor allem in der Entwicklungsphase unverzichtbar.
8: <PB> Ein späterer Schulbeginn ist eine interessante Idee, dessen Umsetzung jedoch nur unter einem Kompromiss möglich ist - dass es unbedingt mit einem Riesenaufwand verbunden ist, was den Schulbusbetrieb oder auch Verpflegung von Schülern angeht, macht sie aber nicht attraktiver.
9: Würde ein Teilverzicht auf sie allerdings eine Verbesserung der Lernbedingungen für Schüler bedeuten, wäre das Gesamtkonzept nicht unwägbar - wichtig ist dabei nur, dass die Schüler eine bessere Schulerfahrung genießen können.

1: _Einleitung
2: _Pro_aus_Lesetext
3: _Pro_aus_Lesetext
4: _Con_aus_Lesetext
5: _Con_aus_Lesetext
6: _Sonstiges
7: _Con_aus_Lesetext
8: _Eigen
9: _Eigen
10: _Sonstiges
11: _Eigen
12: _Sonstiges
13: _Eigen

Table 6: Example invalid output by Llama in the one-shot scenario. Upper part contains the (shortened) prompt containing the input essay, lower part the output by Llama.