# Fine-Grained Content Zone Prediction in German Argumentative Essays Using LLMs

**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 introduce FDE-Arg, a newly compiled dataset of argumentative student essays in German. We use two Llama models of different sizes to label sentence-level content zones both in FDE-Arg and in an existing dataset of source-dependent argumentative essays. We investigate three approaches for improving model performance: a) Incorporating targeted task information into the prompt text; b) few-shot prompting with up to 10 examples selected on the basis of similarity with the target instance; and c) parameter-efficient fine-tuning. We observe that both incorporating additional information in the prompts and similarity-based few-shot prompting have produced highly promising performance gains over the baseline.

## 1 Introduction

Given that written argumentation is a central skill set taught in secondary schools, using the latest NLP applications to support the teaching and analysis of argumentation in student writing is highly desirable. One central subtask is automatically recognising what role individual text units play with respect to the overall text function and extracting those that are relevant to argumentation. As such, the task is related to argumentative zoning as established by Teufel et al. (1999), which targets a different text genre, viz. scientific writing, and analyses argumentative and rhetorical structure in academic papers. Applied to the educational domain, such an analysis of the functional components of a student's argumentative text can contribute to a better understanding of the student's argumentation strategy and form the basis of constructive feedback.

In our present work, we investigate the application of LLMs to this task, which we refer to as "(argumentative) content zone prediction" in German-language argumentative essays. The aim is to automatically assign one out of a set of content zone labels to each sentence in the essay, where the label reveals the functional and argumentative role played by the sentence w.r.t. the essay. To illustrate, Figure 1 shows a fictitious example in English that is loosely adapted from an authentic essay from our German data.

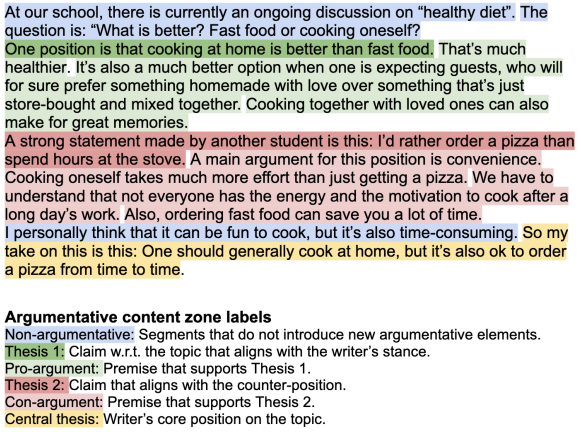


Figure 1: Fictitious example essay with colour-coded annotations, loosely adapted from an authentic German essay from our data.

For this, we use two datasets of argumentative essays in German: The recently released GerTE dataset (Bai and Stede, 2025) featuring source-based essays that are written with reference to reading material, and our novel dataset FDE-Arg, featuring stand-alone essays. Each dataset uses its own set of content zone labels that reflect the respective writing task. For instance, labels in GerTE differentiate between arguments repeated from the reading material and original arguments by the writer (Bai and Stede, 2025), while labels in FDE-Arg differentiate between the writer's position on a discussion topic and the counter-position that the writer concedes.

Bai and Stede (2025) have conducted baseline zone prediction experiments on GerTE using a small Llama model. We take their approach as



*Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications*, pages 405–418

our starting point and investigate ways to improve LLM performance on the argumentative content zone prediction task, using both GerTE and FDE-Arg. Concretely, we experiment with the following three strategies:

- Enhancing the prompt text with task-specific background information
- Few-shot prompting with selection of example instances based on semantic similarity
- Parameter-efficient fine-tuning (PEFT) with QLoRA (Dettmers et al., 2023)

Our overall contributions are two-fold: First, We highlight a novel dataset of German argumentative essays, annotated with a fine-grained set of content zone labels. Second, our experiments reveal that both the inclusion of targeted information in the prompt and few-shot prompting are highly promising directions.

Our code and our resources for FDE-Arg, including the annotation guide as well as annotated datasets, are made publicly available.[1]

## 2 Related Work

Scoring and analysing argumentative essays by students of different age groups is a well-established area of interest in educational NLP: Argumentative essays are featured in well-known datasets such as ASAP-AES[2] and ETS TOEFL11 (Blanchard et al., 2013), which have been widely used in automated essay scoring research (Bai and Stede, 2023). Moreover, the Argument Annotated Essays Corpus (Stab and Gurevych, 2017) and the PERSUADE Corpus (Crossley et al., 2022, 2024) are two prominent examples of English-language learner/student essays which have been annotated with argument mining (AM) information, such as claims and premises, and which therefore allow for the computational analysis of students' argumentation strategies.

With respect to German, DARIUS (Schaller et al., 2024b) is a 4,500-sample corpus of argumentative essays that are written by secondary school students and labelled with multiple layers of argumentation-related annotations, including coarse-grained content zones and major claims. Schaller et al. (2024a) have investigated automatically predicting the labels in DARIUS using both supervised models and an LLM, although their focus is on evaluating the fairness of different models while performing prediction. Stahl et al. (2024) provide a comparable, 1320-sample dataset of student essays with AM labels on multiple levels of granularity. They have also conducted baseline label prediction experiments using a BERT-based model. On the level of university education, Wambsganss et al. (2020a,b) have compiled a 1000-sample corpus of persuasive essays by business administration students and have used it to develop a tool for analysing the argumentation structure in the essays and providing feedback on them.

While to our knowledge LLMs have not yet been extensively applied to these datasets, they have been increasingly applied to the field of AM: Recent work has looked into using LLMs in different prompt scenarios to perform canonical subtasks of AM, such as argument component type recognition and classification (Chen et al., 2024) and argument relation extraction (Gorur et al., 2025). Fine-tuning smaller LLMs using PEFT have also shown to be successful for similar core AM tasks (Kawarada et al., 2024; Cabessa et al., 2025).

## 3 Data

We study argumentative content zone prediction with LLMs, applied to two German corpora with source-dependent and stand-alone essays.

### 3.1 GerTE

GerTE (Bai and Stede, 2025) is a corpus consisting of 117 essays on three topics. The essays have been collected in the context of a source-based writing exercise where the writer first reads a news article on one of three possible topics, then writes an essay that presents the topic, discusses the arguments put forward in the source article and finally concludes with a stance of their own. While the writing task is a common exercise in German secondary schools, GerTE was not collected from authentic school students due to legal concerns. Instead, the authors recruited crowd workers who took on the role of students and composed the essays.

Essays in GerTE have a mean length of approximately 270 words. After automatic sentence-segmentation, the corpus is annotated on the sentence level with functional content zones, where each zone describes a sentence's functional role to

[1] https://github.com/discourse-lab/FairDebArgMining

[2] https://www.kaggle.com/c/asap-aes

the essay. GerTE uses an inventory consisting of the following five class labels:

- **info_intro**: Introductory sentences that present the topic and/or give publication-related information on the source article.
- **article_pro**: Sentences denoting pro-arguments from the source article
- **article_con**: Sentences denoting contra-arguments from the source article
- **own**: Sentences denoting the writer's own position and arguments
- **other**: Sentences that do not fit into the above four classes

The full corpus consists of 1713 sentences from the 117 essays, with each sentence being labelled with exactly one of the five content zones. Label distribution is unbalanced, with "own" being the most frequent and "other" being the least frequent class. The exact distribution is provided in Table 1.

| Label | own | a_pro | info_intro | a_con | other |
|---|---|---|---|---|---|
| Count | 551 | 460 | 281 | 243 | 178 |

Table 1: Class-specific label counts in GerTE

### 3.2 FDE-Arg

Our new dataset comprises stand-alone essays written by students in the 9th grade of various German secondary schools. It was originally collected by education scientists in the context of the *Fair Debating and Written Argumentation* (FDE) project, which aims to investigate and improve the written argumentation skills of 9th-grade students in Germany (Giera et al., 2025a,b). The students were asked to compose pro-and-contra argumentations that should address both sides of a discussion topic, weigh the arguments and draw a conclusion with their own position. Each essay deals with one of four possible, easily accessible topics such as "Is fast food better than cooking at home?" or "Should one undertake volunteer work?". The full FDE dataset comprises 1,061 essays. On this basis, Bai et al. (2026) have developed a scheme for annotating the essays with argumentative content zones[3], which they have applied to and validated on 50 samples. Adopting their scheme, we have annotated another 50 essays, which yields a total set of 100 essays labelled with argumentative content zones. We refer to this 100-sample dataset as *FDE-Arg*.

On the essay level, each text is assigned an "overall argumentative constellation", with "**decided**" denoting that the author argues in favour of a particular stance in the discussion, with or without discussing the opposing stance; "**undecided**" denotes that the author describes both sides of the discussion and refrains from taking a position of their own. Moreover, the essays have been manually segmented into sentences or phrases that form argumentative units,[4] and each segment is labelled with one of the following six labels that describe the segment's argumentative function in the essay:

- **central thesis (cth)**: Segment describing the author's central stance on the topic. If the author subscribes to one side of the discussion without addressing the counter-position, this could be the only thesis/claim in the essay. In essays with the "undecided" overall constellation, "cth" can also be an explicit statement that the author cannot make up their mind.
- **thesis 1 (th1) / thesis 2 (th2)**: Segments describing the two opposing positions on the discussion topic if both are present. In an essay with the "decided" constellation, where the author, however, also acknowledges the counter-position, "th1" denotes the position of the author and is therefore in line with the author's central stance ("cth"), whereas "th2" denotes the counter-position.
- **pro argument (pro) / con argument (con)**: Segment describing arguments that support the theses. Those supporting "th1" and/or "cth" are denoted as "pro arguments" and those supporting "th2" as "con arguments".
- **non argumentative (n-a):** Segment that is not considered argumentative. This includes sentences that serve to structure the essay or that present background information and should be given to all segments not covered by the above labels.

[3]However, they refer to them as *argument component types*, following argumentation mining terminology.

[4]Their rule for manual segmentation is that complex sentences are split into smaller segments if they individually play a role to the argumentative structure of the essay. Those that are not relevant to argumentation, e.g. because they only give background information or serve to structure the essay, are left as they are. We refer to Bai et al. (2026) and their annotation guide for details.

We refer to Bai et al. (2026) for details on the annotation process for FDE-Arg, including theoretical and practical motivations for the scheme used. Moreover, they also present an inter-annotator agreement study based on 30 essays, which has achieved an overall Cohen's $\kappa$ of 0.79.

On average, essays in FDE-Arg are longer than those in GerTE, with a mean word count of 345 per essay. The 100 essays comprise 2,595 labelled segments in total; the label distribution is shown in Table 2. It is unsurprising that there are far fewer segments labelled as "cth", "th1" and "th2" than segments labelled as "pro" and "con" since a single position or claim can be supported by multiple arguments.

| Label | pro | con | n-a | cth | th1 | th2 |
|---|---|---|---|---|---|---|
| Count | 911 | 697 | 688 | 106 | 97 | 96 |

Table 2: Class-specific label counts in FDE-Arg

To summarise, Table 3 captures and contrasts central characteristics of GerTE and FDE-Arg.

## 4 Methods

We use two open-source, instruction fine-tuned LLama 3 models of different sizes: The smaller "Llama-3.1-8B-Instruct"[5] with 8 billion parameters (hereafter referred to as "Llama-8B") and the larger "Llama-3.3-70B-Instruct"[6] with 70 billion parameters (hereafter "Llama-70B"). While newer, more powerful LLMs are a potential way to achieve better prediction results, our focus is on different strategies to enhance model performance while keeping the LLM models themselves fixed. Each of our three strategies is described in detail in the sections to come. Our experiments used the Huggingface framework (Wolf et al., 2020) and Groq API[7].

### 4.1 Baselines

For GerTE, Bai and Stede (2025) have conducted first LLM-based content zone prediction experiments using Llama-8B. They experimented with a zero-shot prompt setting, with and without inclusion of the source article text in the prompt, as well as a one-shot setting using a manually selected example essay for in-context learning. They report their best result as a weighted-average F1 of **0.661**. In contrast, their best-performing ensemble model using BERT and SVM achieved an F1 of 0.713. We treat the prompting approach by Bai and Stede (2025) as our baseline setting on GerTE.

Bai et al. (2026) have also conducted a first set of LLM-experiments, albeit on the initial 50-sample portion of FDE-Arg which they had annotated. Their base prompt text is similar to that by Bai and Stede (2025): It asks the LLM to adopt the role of a 9th-grade teacher, explains the meaning of each label, supplies the segmented essay and asks the model to output a sequence of labels of the same length. Instead of the three-letter labels such as "cth" or "pro" in the dataset, for the prompt text, the zone labels are mapped to interpretable counterparts such as "Zentrale_These" ("central thesis") and "Pro_Argument" ("pro argument"), respectively. The exact German-language prompt text is provided in the Appendix. Using both Llama-8B and Llama-70B, Bai et al. (2026) evaluated zero-shot prompting and two-shot prompting with two randomly chosen example essays, one with a "decided" and one with an "undecided" overall constellation. We adopt these two prompt settings as our baseline on FDE-Arg.

### 4.2 Additional Information in Prompt Text

As the most simple strategy, we experimented with enhancing the baseline prompt for either dataset with additional information that addresses specific challenges in the respective dataset.

#### 4.2.1 GerTE: Summary of Source Text Content

The baseline by Bai and Stede (2025) has already explored incorporating the source article text into the LLM prompt. This has indeed benefitted the LLM's prediction, although the resulting performance was still far below that of supervised models. According to the authors' error analysis, a significant source of errors is confusion between pro and con arguments from the source article as well as their distinction from the writer's own arguments. Since a solid understanding of the main arguments from the source text is essential to avoiding such confusion, we manually extracted and summarised the main arguments from each source article in concise sentences[8] and included them in the system

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[7] https://groq.com/

[8] These summarised main arguments are in fact what has been referred to as "reference snippets" in Bai and Stede (2022). In a preparatory step to composing the essays in

| | GerTE | FDE-Arg |
|---|---|---|
| Essay type | source-dependent | stand-alone |
| Mean word count per essay | 270 | 345 |
| Label set | own, article_pro, article_con, info_intro, other | cth, th1, th2, pro, con, n-a |
| Annotation unit | sentence (automatic sentence-segmentation) | argumentative segment (manual segmentation) |
| Total number of essays | 117 | 100 |
| Total number of annotated units | 1,713 | 2,595 |

Table 3: Central characteristics of GerTE and FDE-Arg

prompt alongside the source article itself. A gloss of our addition is thus: *The following are the arguments that are brought up in the text. The pro arguments are these [LIST OF PRO-ARGUMENTS]. The con arguments are these [LIST OF CONTRA-ARGUMENTS].* The full system prompt for an essay on the topic of using social media in class is provided in the Appendix, including our formulation of the article's main arguments.

#### 4.2.2 FDE-Arg: Instructions from Annotation Guide

The annotation guide for FDE-Arg provides detailed information and annotation strategies on the dataset. Therefore, we aimed to improve LLM performance by adding relevant content from the annotation guide to the prompt text, experimenting with two sets of information in particular:

First, various authors of the essays conclude their essay by stating or repeating their overall view on the discussion topic. According to the annotation scheme, such a statement should be annotated as "central thesis (cth)" only if no previous segment has been annotated as such. Otherwise, it is considered a conclusion and thus an essay structuring element that is annotated as "non-argumentative (n-a)". To reduce confusion between "cth" and "n-a", we added a formulation of this annotation rule to the prompt. The exact formulation of this addition is given in the Appendix.

Second, for further analysis of an essay's argumentation structure, it is particularly crucial that the theses, particularly the central thesis, are correctly extracted, and human annotators are instructed by the annotation guide to prioritise the labels for theses over those for arguments in cases of doubt. To encourage the same behaviour from LLMs and to target better prediction of the thesis labels, we added a short instruction based on the recommended annotation steps from the annotation guide. A gloss of our addition is as follows: *When assigning zone labels to segments you can proceed as follows: 1. Find the central message of the essay ("cth") ... 2. Find further theses ("th1" / "th2") ... 3. Find the arguments ("pro" / "con") ... 4. Mark the remaining, unassigned segments ("n-a") ....* We refer to the Appendix again for the full addition.

### 4.3 Similarity-Based Few-Shot Prompting

In contrast to the one-shot and two-shot prompts in the baseline settings (see above), we experimented with few-shot prompting using up to 10 example essays. Moreover, previous work on example selection for in-context learning have shown that example instances in few-shot prompting are especially effective when they display high levels of similarity with the target instance (Liu et al., 2022; Ajjour and Wachsmuth, 2025). We therefore adopted a similarity-based method to individually and automatically select a set of example essays for each target essay, as described in the following.

Although we did not *train* the models in any sense, we divided our data into a training and a test set, of which the training set served as the pool of essays from which examples could be selected. For a given test essay $t$ at inference time, we first extracted all training essays that address the same topic as $t$, treating them as "candidate examples" $C$. Next, we ranked all essays in $C$ based on their semantic similarity with $t$ and sorted them in descending order. From this ranked list of candidate essays, we then selected the top $k$ essays, where $k$ is the number of examples we used in the few-shot prompting scheme.

Semantic similarity between a given candidate essay $c \in C$ and $t$ was computed using SBERT (Reimers and Gurevych, 2019), specifically the pre-trained model "sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2"[9], and cosine simi-

GerTE, writers were in fact asked to write out the main arguments from the source article. These summarised arguments we include in the prompts have originally been manually compiled to serve as reference answers for this preparatory task.

[9] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

larity. For this calculation, we split both $c$ and $t$ into three chunks, respecting sentence boundaries, and recorded the similarity between the index-aligned pairs of chunks. This produced three similarity values for each essay pair. The final similarity value between the essay pair was computed as a weighted average of the three chunk-level similarity scores. Our similarity-based example ranking process is illustrated in Figure 2.

The motivation for splitting each candidate and target essay into three chunks is two-fold: First, entire essays would exceed the 128-token sequence length limit of the SBERT model used, which would subject them to truncation. Second, by taking a weighted average of chunk-level similarities, we could control the weight put on the similarity in specific areas of essays. For FDE-Arg essays, we assigned equal weight to all three chunks. For GerTE, in contrast, we assigned less weight to the similarity score of the top chunk pair since most essays in GerTE start by providing meta data on the source article. We considered similarity in this area to be less relevant.

We performed our experiments on both datasets using cross-validation, with $k \in \{1, 2, 4, 6, 8\}$ for GerTE and $k \in \{1, 2, 4, 6, 8, 10\}$ for FDE-Arg. If $k$ is larger than the number of candidate examples available[10], all candidates were selected and used.

### 4.4 Parameter-Efficient Fine-tuning

As an alternative to improving the prompt fed to the LLMs, we also experimented with parameter-efficient fine-tuning (PEFT) using QLoRA (Dettmers et al., 2023) with 4-bit quantization. Due to computational costs, this approach was only applied to the smaller model Llama-8B. We again divided our data into training and test partitions through 5-fold cross-validation and fine-tuned the model on the training partition. For either dataset, the model was fine-tuned on the zero-shot baseline prompt text, paired with the target output, which could be derived from the gold-standard zone labels in the datasets. The approach was implemented using the PEFT functionalities that are integrated into the Huggingface framework. All hyperparameters used are provided in the Appendix.

[10]This can happen if the training portion in a given fold does not contain sufficient essays on the same topic as the target essay.

## 5 Results

For both datasets, we adopted a regex-based method of extracting labels from the LLM's output. Where the model's output used made-up zone labels or where it did not produce the same number of labels as the number of input segments, we considered the output invalid and excluded it from evaluation.

Based on all test essays that yielded valid output, we evaluated model performance using standard multi-class classification metrics, viz. accuracy and average F1. For comparability, we follow Bai and Stede (2025) and used the *weighted* average F1 for GerTE; for FDE-Arg, we saw no justification for assigning frequency-dependent weights to the different content zones and therefore used the *macro* average F1.

Our experiment results are summarised in Table 4 for GerTE[11] and in Table 5 for FDE-Arg[12]. Where cross-validation was used, the metrics are as averaged across all folds. The best-performance on either dataset by either model is printed in bold.

Our results reveal the following general trends: First, in line with expectations, the larger Llama model with 70 billion parameters consistently outperforms the smaller one. On GerTE, to which the larger model has not yet been applied, it achieves a weighted F1 of 0.79 under the best experimental condition (similarity-based six-shot prompting). This by far beats the hitherto highest F1 of 0.713, which is the performance of the best supervised model reported by Bai and Stede (2025), while requiring only six labelled essays. Moreover, with the same six-shot prompting scheme, the smaller Llama-8B model in fact also surpasses said supervised model, achieving an F1 of 0.73.

Second, confirming Bai et al. (2026)'s observations, zone extraction on FDE-Arg proves to be highly challenging, particularly for the smaller Llama-8B model. In fact, it produced invalid outputs for one-third to half of the test essays, and the labelling results on those that did receive valid outputs were consistently poor. Upon manual examination of the invalid outputs, we found that nearly all of them are cases of made-up labels that were simply not the ones that the model had been

[11]Here, the results for the baselines using Llama-8B are taken from Bai and Stede (2025)'s original paper.

[12]Note that the baseline performance cannot be directly compared to the results in Bai et al. (2026) since their experiments only used 50 essays rather than our 100-sample FDE-Arg.

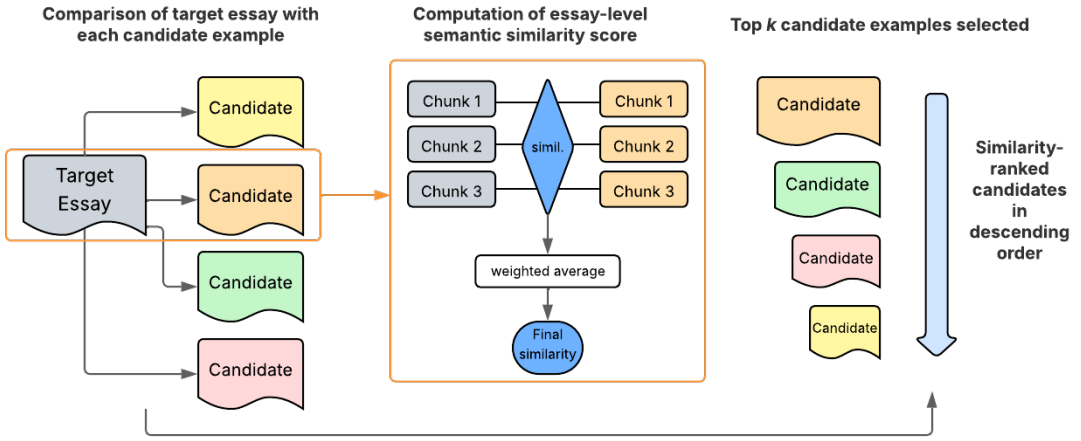


Figure 2: Ranking of candidate example essays for a given target essay based on semantic similarity.

| | Llama-8B | | | Llama-70B | | |
|---|---|---|---|---|---|---|
| | # invalid | Acc | F1 (w-avg) | # invalid | Acc | F1 (w-avg) |
| zero-shot (baseline) | 12 | 0.654 | 0.641 | 1 | 0.736 | 0.731 |
| zero + article (baseline) | 8 | 0.673 | 0.661 | 0 | 0.755 | 0.745 |
| one-shot (baseline) | 10 | 0.661 | 0.649 | 0 | 0.751 | 0.743 |
| zero + prompt addition (article + arguments) | 8 | 0.686 | 0.672 | 0 | 0.776 | 0.77 |
| one + prompt addition (article + arguments) | 7 | 0.687 | 0.680 | 0 | 0.78 | 0.774 |
| Similarity-based few-shot prompting | | | | | | |
| $k = 1$ | 6 | 0.703 | 0.686 | 1 | 0.762 | 0.748 |
| $k = 2$ | 7 | 0.707 | 0.691 | 0 | 0.766 | 0.751 |
| $k = 4$ | 17 | 0.716 | 0.702 | 0 | 0.772 | 0.755 |
| $k = 6$ | **9** | **0.742** | **0.73** | **0** | **0.8** | **0.79** |
| $k = 8$ | 12 | 0.732 | 0.721 | 1 | 0.782 | 0.762 |
| PEFT | 2 | 0.635 | 0.638 | - | - | - |

Table 4: Content zone prediction results on GerTE in terms of accuracy (Acc) and weighted-average F1 (F1, w-avg). "# invalid" refers to the number of test essays which were excluded from evaluation due to invalid output.

instructed to use. The same behaviour of generating illegal labels is also observed with Llama-8B on GerTE, though much less frequently (see Table 4). Overall, the F1 scores achieved by Llama-8B should be interpreted with the number of invalid outputs in mind and are arguably less conclusive.

Finally, contrary to expectations, our fine-tuning experiments using QLoRA on Llama-8B did not bring about any improvements in terms of the chosen metrics. Our fine-tuned models performed near or even below the base models. We do observe, nonetheless, that on both datasets, the fine-tuning did lead the model to produce far fewer instances of invalid output. This is particularly true for FDE-Arg (see Table 5).

To illustrate common challenges in our task, Figure 3 shows the truth-normalised confusion matrices (values within rows add up to 1) for the best-performing model, which, for either dataset, is Llama-70B in a similarity-based few-shot prompt scenario. The values are as averaged across all folds in cross-validation. For GerTE, "other" stands out as the most difficult class to recognise. As pointed out by Bai and Stede (2025), this is expected since it is both the least frequent label and a fuzzily defined fall-back class for all sentences that do not fit elsewhere. For FDE-Arg, confusion between "th1" and "th2" and, concurrently, between "pro" and "con" form the main challenges.

## 6 Discussion and Further Analysis

The focus of our experiments lies in exploring the three aforesaid strategies for improving LLM performance on the content zone prediction task. Based on our results, similarity-based few-shot prompting with approximately 6 to 8 labelled essays is particularly promising, showing the best prediction results on both datasets. This raises the question whether this success can be attributed to the sheer use of more example essays or to the

| | Llama-8B | | | Llama-70B | | |
|---|---|---|---|---|---|---|
| | # invalid | Acc | F1 (m-avg) | # invalid | Acc | F1 (m-avg) |
| zero-shot (baseline) | 45 | 0.51 | 0.380 | 1 | 0.576 | 0.490 |
| two-shot (baseline) | 41 | 0.628 | 0.514 | 1 | 0.745 | 0.632 |
| zero + "cth" clarification | 48 | 0.525 | 0.397 | 3 | 0.618 | 0.529 |
| two + "cth" clarification | 48 | 0.593 | 0.448 | 4 | 0.684 | 0.640 |
| zero + recommended prediction steps | 44 | 0.421 | 0.329 | 1 | 0.588 | 0.530 |
| two + recommended prediction steps | 41 | 0.607 | 0.482 | 1 | 0.686 | 0.649 |
| Similarity-based few-shot prompting | | | | | | |
| $k = 1$ | 39 | 0.549 | 0.446 | 0 | 0.641 | 0.585 |
| $k = 2$ | 46 | 0.567 | 0.476 | 1 | 0.708 | 0.66 |
| $k = 4$ | 43 | 0.599 | 0.471 | 1 | 0.749 | 0.698 |
| $k = 6$ | **38** | **0.612** | **0.519** | 2 | 0.728 | 0.685 |
| $k = 8$ | 38 | 0.602 | 0.499 | **0** | **0.773** | **0.750** |
| $k = 10$ | 36 | 0.597 | 0.504 | 1 | 0.76 | 0.729 |
| PEFT | 6 | 0.547 | 0.408 | - | - | - |

Table 5: Content zone prediction results on FDE-Arg in terms of accuracy (Acc) and macro-average F1 (F1, m-avg). "# invalid" refers to the number of test essays which were excluded from evaluation due to invalid output.

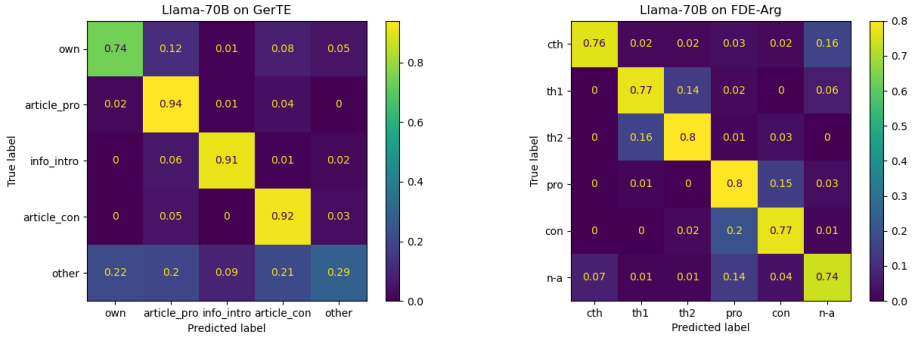


Figure 3: Truth-normalised confusion matrices for the best-performing experimental setting on either dataset.

semantic similarity between example and target essays. To further examine this, we also experimented with $k$-shot prompting with *random* example selection. In this case, $k$ examples are simply chosen at random from the set of candidate essays, viz. training essays that address the same topic as the target essay.

Appendix C provides the detailed F1 scores achieved by the models using randomly selected example essays, once again as averaged across all folds in cross-validation. Figure 4 graphically compares the F1 scores[13] obtained in the similarity-based vs. the random few-shot prompting scheme. We have left out the performance of Llama-8B on FDE-Arg, where the large amounts of invalid outputs render the results inconclusive.

Overall, in both similarity-based and random few-shot prompting, an improvement of model performance can be observed with increasing numbers of examples $k$ until it plateaus at approximately $k = 6$ or $k = 8$. While choosing examples based on semantic similarity with target samples does not outperform choosing them randomly across all values for $k$, the similarity-based scheme is superior in the majority of the cases, especially with the application of Llama-70B to FDE-Arg. Moreover, across the board it is the similarity-based prompt setting that produces the best performance for a given model-dataset combination.

While not as effective as few-shot prompting, our experiments have also shown that adding task-relevant information to the prompt text can indeed benefit LLMs' zone prediction performance. This is especially attractive since this strategy is highly cost-efficient, requiring neither significantly more computational resources (unlike PEFT) nor the availability of labelled training data. As shown by Table 4 and Table 5, in all of our experiments, the inclusion of our chosen additional information in the prompt text produced performance improvement compared to the baseline settings. The only exception here is that of Llama-8B on FDE-Arg, which, again, is less conclusive due to the abundance of invalid output. With respect to GerTE,

[13] Once again, *weighted*-average F1 for predictions on GerTE, *macro*-average F1 for those on FDE-Arg.

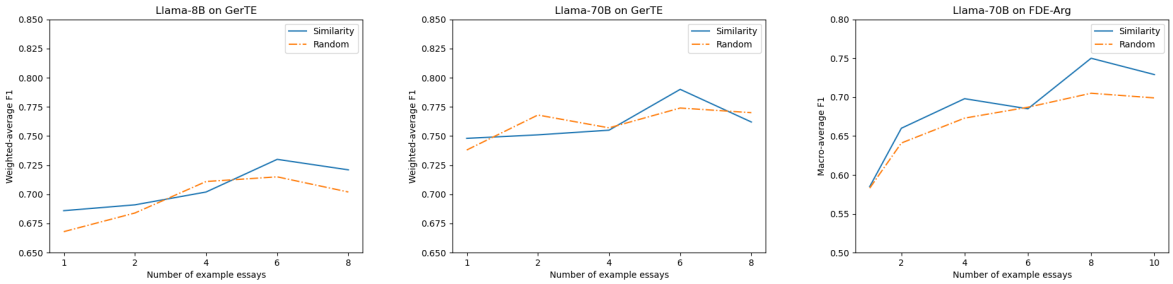


Figure 4: Graph representation of the F1 scores in the similarity-based (solid blue line) vs. the random (dashed orange line) $k$-shot prompting scheme.

the inclusion of our summary of the source article arguments primarily aims at improving the model's performance on the classes "article_pro", "article_con" and "own". Similarly, the annotation guide-inspired additions to the prompt text for FDE-Arg have been designed to target correct prediction of the classes "cth", "th1" and "th2". Table 6 and Table 7 respectively show the class-specific F1 values as achieved by Llama-70B for the targeted classes in either dataset, contrasting the baselines with the settings including additional information in the prompt. It can be observed that the targeted classes have benefitted from the additions almost across the board.

| Label | a_pro | a_con | own |
|---|---|---|---|
| zero (baseline) | 0.788 | 0.701 | 0.766 |
| zero + article + argument | 0.869 | 0.777 | 0.772 |
| one (baseline) | 0.804 | 0.694 | 0.793 |
| one + article + arguments | 0.867 | 0.761 | 0.799 |

Table 6: Llama-70B on GerTE: Class-specific F1 for "article_pro" ("a_pro"), "article_con" ("a_con") and "own" in selected experiment conditions.

| Label | cth | th1 | th2 |
|---|---|---|---|
| zero (baseline) | 0.454 | 0.26 | 0.392 |
| zero + "cth" clarification | 0.496 | 0.27 | 0.444 |
| zero + prediction steps | 0.476 | 0.333 | 0.536 |
| two (baseline) | 0.561 | 0.522 | 0.618 |
| two + "cth" clarification | 0.554 | 0.532 | 0.639 |
| two + prediction steps | 0.576 | 0.557 | 0.657 |

Table 7: Llama-70B on FDE-Arg: Class-specific F1 for "cth", "th1" and "th2" in selected experiment conditions.

Finally, as remarked in the previous section, fine-tuning using PEFT and QLoRA has not proved useful in our experiments. We do note, however, that both datasets we used are small for the purpose of model training, and that we have not performed any hyperparameter tuning. We therefore consider this result to be tentative and in need of further experimentation.

## 7 Conclusion

We have introduced the novel German dataset FDE-Arg, which consists of 100 stand-alone argumentative essays by secondary school students and are annotated with a fine-grained set of argumentative content zones as developed by Bai et al. (2026). Using both FDE-Arg and source-dependent essays from the GerTE dataset, we applied two differently-sized Llama models to the task of predicting the argumentative content zones. While promising results could be obtained from the larger model, the smaller model struggled with the task, especially with the more challenging FDE-Arg dataset.

We further focused on three strategies for boosting model performance and found both the simple incorporation of additional, task-related information into the prompt text and few-shot prompting with approximately 6 to 8 labelled example instances to be effective. Moreover, choosing the examples based on semantic similarity with the target instance seems to be particularly promising. We could not observe any performance improvement through our current implementation of the fine-tuning approach using PEFT . However, FDE-Arg is gradually being expanded, and argumentative zone labels are being applied to a growing portion of the original FDE dataset. Therefore, we plan to revisit PEFT-based strategies in the future using significantly more labelled data and newer LLMs. Moreover, given the success of our similarity-based few-shot prompting scheme, we also plan to explore alternative ways to capture similarity between target and example instances in order to improve example selection for in-context learning.

## Limitations

We acknowledge the following limitations in relation to our work: First, both datasets we used are rather small. GerTE contains 117 essays, while FDE-Arg contains 100. However, as mentioned in the main text, FDE-Arg is growing, and the original FDE dataset consists of 1,061 essays, which we plan to fully annotate with argumentative content zones within the next months. We are therefore confident that we will soon be able to elaborate on our experiments with more labelled data. Second, our experiments with PEFT have only been a first step, and we did not perform hyperparameter tuning. We plan to look more into fine-tuning-based approaches using LLMs once more training data from FDE-Arg are available. Finally, for now, with Llama 3 we have chosen a well-known but comparatively old LLM since our focus does not lie in comparisons between different LLMs. We do plan to repeat our experiments with more up-to-date LLMs in the future.

## Acknowledgments

We thank the anonymous reviewers for their highly helpful comments and suggestions on the first draft of this paper. We are also grateful to Jun. Prof. Winnie-Karen Giera and her team for building the original FDE corpus and cooperating with us. Kemal Afzal and Dietmar Benndorf have made significant contributions to the annotation scheme for FDE-Arg. Part of the work reported here is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 567969163.

## References

Yamen Ajjour and Henning Wachsmuth. 2025. Exploring LLM Priming Strategies for Few-Shot Stance Classification. In *Proceedings of the 12th Argument mining Workshop*, pages 11–23, Vienna, Austria. Association for Computational Linguistics.

Xiaoyu Bai, Kemal Afzal, Dietmar Benndorf, Lucas Deutzmann, Winnie-Karen Giera, Eric Graßnick, and Manfred Stede. 2026. From newspapers to classrooms: Adapting an annotation scheme and automatic classifiers to mixed-quality argumentative school essays. Submitted.

Xiaoyu Bai and Manfred Stede. 2022. Argument Similarity Assessment in German for Intelligent Tutoring: Crowdsourced Dataset and First Experiments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2177–2187, Marseille, France. European Language Resources Association.

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.

Xiaoyu Bai and Manfred Stede. 2025. Predicting Functional Content Zones in German Source-Dependent Argumentative Essays: Experiments on a Novel Dataset. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 66–79, Hannover, Germany. HsH Applied Academics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument Mining with Fine-Tuned Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the Potential of Large Language Models in Computational Argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

W. K. Giera, M. Stede, L. Deutzmann, and E. Graßnick. 2025a. Exploring the Power of Persuasion in Written Argumentation: A Mixed-Methods Pilot Study (QASA). *Journal of Applied Language Learning*, 2(2):1–10.

Winnie-Karen Giera, Lucas Deutzmann, and Subhan Sheikh Muhammad. 2025b. Merging oral and written argumentation: Supporting student writing through debate and srsd in inclusive classrooms. *Education Sciences*, 15(11):1471.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can Large Language Models perform Relation-based Argument Mining? In *Proceedings of the 31st International Conference on Computational Linguistics*,

pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Masayuki Kawarada, Tsutomu Hirao, Wataru Uchida, and Masaaki Nagata. 2024. Argument Mining as a Text-to-Text Generation Task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2002–2014, St. Julian's, Malta. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. 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, 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 Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024b. DARIUS: A Comprehensive Learner Corpus for Argument Mining in German-Language Essays.

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.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh Edinburgh, Scotland.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020a. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery, New York, NY, USA.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020b. A Corpus for Argumentative Writing Support in German. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 856–869, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

## A LLM Prompts

The following shows a baseline prompt used on FDE-Arg, featuring an authentic target essay from the dataset.

*System prompt*:

Du bist ein Deutschlehrer der 9. Klasse und analysierst argumentative Aufsätze deiner Schüler. Die Schüler schreiben den Aufsatz mit Bezug auf eine vorgegebene Streitfrage. Du überprüfst, wie der Aufsatz argumentativ aufgebaut ist.

*User prompt*:

Im Folgenden liegt ein Aufsatz vor. Der Aufsatz wurde in kleineren Einheiten wie Sätzen oder Phrasen unterteilt, die im folgenden zur Vereinfachung als 'Sätze' bezeichnet werden. Die Sätze sind nummeriert. Das Format dabei is 'Satznummer: Satz'.
Entscheide zunächst, welche argumentative Gesamtkonstellation vorliegt:
'Entschieden' bedeutet: Der Aufsatz argumentiert für eine bestimmte Position des Autors.
'Unentschieden' bedeutet: Der Aufsatz bringt gleichberechtigt Argumente für beide Seiten der Streitfrage hervor. Der Autor entscheidet sich nicht für eine Position.
Analysiere die Funktionen der einzelnen Sätze. Dabei gilt eine grundlegende Unterscheidung zwischen der Funktion der These und der des Arguments: Eine These ist eine argumentative Position, die jemand einnehmen kann. Ein Argument ist ein Grund, der eine bestimmte These stützt oder angreift.
Jeder Satz im Aufsatz soll einer der folgenden 6 Funktionen zugeordnet werden:
'Zentrale_These' bedeutet: Der Satz beschreibt in einem entschiedenen Aufsatz die Kernposition

des Autors, oder er drückt in einem unentschiedenen Aufsatz explizit aus, dass der Autor sich nicht für eine Position entscheiden kann.
'These_1' bedeutet: Der Satz beschreibt in einem entschiedenen Aufsatz die Position, die mit der zentralen These übereinstimmt, oder er beschreibt in einem unentschiedenen Aufsatz die im Text erstgenannte Position.
'These_2' bedeutet: Der Satz beschreibt in einem entschiedenen Aufsatz die Position gegen die der zentralen These, oder er beschreibt in einem unentschiedenen Aufsatz die im Text später genannte Position.
'Pro_Argument' bedeutet: Der Satz beschreibt ein Argument, das These 1, also die Position des Autors in einem entschiedenen Aufsatz, stützt und bestärkt.
'Con_Argument' bedeutet: Der Satz beschreibt ein Argument, das These 2, also die Position gegen die des Autors in einem entschiedenen Aufsatz, stützt und bestärkt.
'Nicht_Argumentativ' bedeutet: Der Satz hat keine argumentative Funktion. Stattdessen gibt er zum Beispiel Hintergrundinformationen zum Thema, Anekdoten des Autors oder dient der Gliederung und der abschließendenZusammenfassung des Aufsatzes.
Werte den Aufsatz aus und ordne jeden Satz einer der genannten 6 Funktionen zu. Das Format des Outputs soll ausschließlich wie folgt sein:
'Gesamtkonstellation
Satznummer: Funktion
Satznummer: Funktion'
, zum Beispiel
'Entschieden
1: Zentrale_These
2: Nicht_Argumentativ
3: Pro_Argument
4: These_1'.
Im Output sollen es neben der Gesamtkonstellation genauso viele Satznummer-Funktion-Paare geben wie es Sätze im Aufsatz gibt.

Schüleraufsatz

1: Was ist nun besser, Fast Food oder selbsgekochtes, gesundes Essen?
2: Immer öfter findet eine Diskussion zwischen den beiden Seiten statt, in der über die bessere Art von Essen debattiert, doch was ist nun besser?
3: Heutzutage konsumieren Kinder oder Teenager immer mehr Fast Food,
4: es ist schnell, günstig und lecker.
5: Dies stellt eine große Erleichterung für sie dar, wenn sie zum Beispiel nach einem stressigen Schultag schnell in ein Fast Food Restaurant gehen können und innerhalb kürzester Zeit ihr Essen vor sich stehen haben.
6: Jedoch kommen auch Argumente ans Licht die definitiv gegen eine solche Ernährung sprechen und eine gesündere Ernährung befürworten.
7: Aus diesem Anlass wird in dieser Erörterung über die postiven Seiten beider Essensarten, aber auch über ihre Probleme gesprochen.
8: Selbst gekochtes Essen ist viel gesünder
9: und besser als Fast Food,
10: da es eine bessere Essensqualität hat, das steht nicht zur Frage.
11: Man nimmt weniger Zusatzstoffe und Kalorien zu sich,
12: da man frische Sachen wie Gemüse verwendet.
13: Selbstkochen kann Spaß machen
14: und man findet vielleicht ein neues Hobby am Kochen,
15: doch auch Fast Food hat seine Vorteile.
16: Im Gegensatz zum selbstgekochten gesunden Essen, ist dieses günstiger und schneller,
17: fast immer schmeckt es einem
18: und nicht jeder hat noch Energie um nach einem anstrengendem Tag m Herd zustehen und zu kochen.
19: Fast Food findet man ja auch heutzutage überall, wenn man zum Beispiel in die Stadt geht um zu shoppen oder einkaufen zu gehen.
20: Alle Seiten haben ihre Vorteile aber auch Nachteile vorallem im Thema Gesundheit, wo frisch zubereitetes Essen auf jedenfall besser für den Körper ist doch man kann beide Seiten verstehen.
21: Zusammenfassend kan man sagen, dass Fast Food natürlich deutlich weniger Aufwand macht, zudem günstiger ist, aber auf eine gesunde Ernährung kann man nicht verzichten.
22: Meiner Meinung nach sollte man sich in der Mitte treffen.
23: Selbstgekochte, gesündere Sachen sind zwar besser und tragen zu einer guten Ernährung bei, aber man kann sich ab und zu einen Burger gönnen.
24: Doch übertreiben sollte man es nicht.
25: Zum Schluss lässt sich sagen, dass jeder das essen soll, was er/sie am besten für sich selber hält, aber eine gesunde Ernährung schaded nie.

The following shows a system prompt on GerTE with the added summary of source article arguments (see Section 4.2.1). The user prompt is the same as in the baseline prompt scenario.

Du bist ein Deutschlehrer der 9. Klasse und analysierst Aufsätze deiner Schüler. Du überprüfst dabei, ob die Aufsätze alle zu erwartenden Bestandteile enthält.
Hier ist der Lesetext, auf den der Aufsatz Bezug nimmt.

[FULL SOURCE ARTICLE TEXT]

Hier sind die Argumente, die im Lesetext vorkommen: Die Pro-Argumente sind folgende:
Unterricht ist abwechslungsreicher.
Unterricht bleibt länger im Gedächtnis.
Nutzung von Twitter senkt die Hemmung vor aktiver Beteiligung am Unterricht.
Auch schüchterne Schüler:innen werden ermutigt, ihre Meinungen zu äußern.
Pädagogen raten zur Auseinandersetzung mit dem medialen Wandel.
Es gibt Erfolgsbeispiele für die Nutzung von Twitter im Unterricht, z.B. in den USA.
Angst vor Ablenkung ist nicht begündet, da auch Bücher ablenken können.
Laut Psychologen fördert das Internet die Gesprächskultur.

Die Con-Argumente sind folgende:
Schüler können leicht abgelenkt werden.
Viele Lehrer sind neuen Medienformen gegenüber skeptisch eingestellt.

The following shows the addition to the prompt text for FDE-Arg concerning the distinction between “cth" and “n-a" (see Section 4.2.2).

> Manche Aufsätze enden mit einer knappen Zusammenfassung der Position des Autors, zum Teil etwas umformuliert. Sofern die zentrale These bereits woanders markiert wurde, wird dieser Abschlusssatz als Zusammenfassung des Aufsatzes gewertet und somit als “Nicht_Argumentativ" markiert. Gibt es hingegen an keiner anderen Stelle die zentrale These, so wird dieser Abschlusssatz als “Zentrale_These" markiert.

The following shows the addition to the FDE-Arg prompt text concerning the recommended prediction steps (see Section 4.2.2).

> Bei der Zuordnung kannst du wie folgt vorgehen:
> 1. Finde die Kernaussage des Textes (“Zentrale_These"): Diese steht für sich selbst und kann prinzipiell an jeder Stelle des Textes stehen.
> 2. Finde weitere Thesen (“These_1" / “These_2"): Wenn der Text beide Positionen beleuchtet, unabhängig davon, ob der Autor sich letztendlich für eine Seite entscheidet, werden die beiden Positionen markiert.
> 3. Finde Argumente (“Pro_Argument" / “Con_Argument"): Argumente begründen die Thesen. Dabei können sie die Thesen direkt stützen oder ein anderes Argument und damit indirekt die Thesen stützen.
> 4. Markiere verbleibende Sätze (“Nicht_Argumentativ"): Alle Sätze, die zuvor nicht markiert wurden, haben keine argumentative Funktion.

## B PEFT Hyperparameters

Hyperparameters for the PEFT-experiments on either dataset are given in Table 8. All parameters not mentioned in the table used the default values by HuggingFace.

## C Few-Shot Prompting with Randomly Selected Examples

Table 9 and Table 10 give the experimental results on few-shot prompting using *randomly* selected example essays. All values are as averaged across all folds in cross-validation.

| | **GerTE** | **FDE-Arg** |
|---|---|---|
| learning rate | $5e^{-5}$ | $2e^{-4}$ |
| per_device_train_batch_size | 4 | 4 |
| gradient_accumulation_steps | 2 | 4 |
| maximum training steps | 320 | 300 |
| LoRA-specific hyperparameters | | |
| r | 8 | 16 |
| lora_alpha | 16 | 16 |
| lora_dropout | 0.05 | 0 |
| bias | none | none |
| task_type | CAUSAL_LM | CAUSAL_LM |

Table 8: Hyperparameters used in the PEFT-experiments on either dataset.

| | **Llama-8B** | | | **Llama-70B** | | |
|---|---|---|---|---|---|---|
| | # invalid | Acc | F1 (w-avg) | # invalid | Acc | F1 (w-avg) |
| $k = 1$ | 13 | 0.688 | 0.668 | 1 | 0.755 | 0.738 |
| $k = 2$ | 15 | 0.699 | 0.684 | 0 | 0.784 | 0.768 |
| $k = 4$ | 7 | 0.726 | 0.711 | 0 | 0.773 | 0.757 |
| $k = 6$ | 8 | 725 | 0.715 | 1 | 0.789 | 0.774 |
| $k = 8$ | 12 | 0.72 | 0.702 | 0 | 0.787 | 0.77 |

Table 9: Prediction results on GerTE using the few-shot prompting scheme with *randomly* selected examples.

| | **Llama-8B** | | | **Llama-70B** | | |
|---|---|---|---|---|---|---|
| | # invalid | Acc | F1 (m-avg) | # invalid | Acc | F1 (m-avg) |
| $k = 1$ | 50 | 0.533 | 0.413 | 0 | 0.638 | 0.583 |
| $k = 2$ | 41 | 0.579 | 0.462 | 0 | 0.697 | 0.641 |
| $k = 4$ | 41 | 0.637 | 0.531 | 0 | 0.715 | 0.673 |
| $k = 6$ | 37 | 0.565 | 0.466 | 1 | 0.728 | 0.687 |
| $k = 8$ | 35 | 0.621 | 0.505 | 1 | 0.735 | 0.705 |
| $k = 10$ | 41 | 0.612 | 0.506 | 1 | 0.741 | 0.699 |

Table 10: Prediction results on FDE-Arg using the few-shot prompting scheme with *randomly* selected examples.