Grade Level		Gender		Profile		Language	
Level	**Students**	**Gender**	**Students**	**Profile**	**Students**	**Language**	**Students**
9	423	Female	801	Natural Sciences	414	native	1265
10	346	Male	664	Social Sciences	255	non-native	576
11	547	Diverse	90	Sports	119		
12	404	Missing	284	Linguistics	61		
13	113			Aesthetics	13		
Missing	6			Missing	977		

Table 1: Combined Overview: Grade Level, Gender, Profile, and Language of Students

KFT The Cognitive Abilities Test (*Kognitiver Fähigkeitstest* or KFT) developed by Heller and Perleth (2000), measures students' cognitive abilities through non-verbal figural analogies. These questions evaluate abstract reasoning and the ability to apply logical rules to visual information without linguistic content, making them useful for assessing individuals across different linguistic backgrounds. A typical problem displays a sequence of shapes that follow a certain transformation (e.g., rotation, reflection). The test-taker must identify and apply the same transformation to a new set of figures.

4 Method

In the following section, we describe the experimental setup for our evaluation study.

4.1 Classifiers

We experiment with a diverse set of classifiers to see performance and fairness differences between instances of different model architectures. Our machine learning goal is to predict certain spans in an essay text. For most of these spans, span boundaries align with sentence boundaries.

Major claim annotations always consist of single sentences. The other annotation types, i.e. content zone and position annotations may also span multiple sentences. Only warrant annotations do not necessarily align with sentence boundaries and can consist of segments on the sub-sentence level. Therefore, we make use of both sentence classification and sequence tagging approaches. For sentence classification, we use a Support Vector Machine (SVM) in standard configuration, provided by the scikit-learn python package (Pedregosa et al., 2011) as an instance of shallow learning. The features utilized in the SVM classifier are the TF-IDF vectors of the most frequent 1- to 3-grams. We use a BERT-based [3] sentence classifier as an instance of deep learning and GPT-4 (OpenAI, 2024) to represent generative LLMs. For sequence tagging, we also use the BERT-based classifier and again prompt GPT-4 this time providing the whole essay as input.

[3] dbmdz/bert-base-german-cased

4.2 Data Split

We use a fixed data split of 80% training data and 20 % test data. From the training data, we used a subset of 60% as validation data to find the best epoch for deep learning and for prompt-tuning for generative LLMs in pre-experiments, i.e. the whole training data set was used in the main experiments for training. As we were not interested in the overall best performance but rather in the intrinsic fairness differences between models, we did not further fine-tune any hyperparameters.

4.3 Performance and Fairness Evaluation

The evaluation of our classification results is motivated by the intended use of the classifiers to provide formative feedback to learners in e.g. an online tutoring system. Although it might also be of interest to show the specific location of an argumentative element within a learner essay as feedback, our primary concern for this study is to determine whether certain argumentative elements are present in a text or not. Therefore, we first transform any classifier output into a binary decision on the document level indicating whether (at least one instance of) a certain argumentative element is present in an essay.

In our fairness evaluation, we follow the framework proposed by (Loukina et al., 2019) and their implementation provided within the RSMTool software package (Madnani and Loukina, 2016). More precisely, we compute *overall score accuracy (osa)*, *overall score difference (osd)* and *conditional score difference (csd)*, where the first looks at squared errors $(S - H)^2$ and the latter two at actual errors $S - H$. In every case, a linear regression is fit with the error being the dependent variable and the

Label	Model	All	Grades	Gender	Profile	School	Languages	KFT
Introduction	Shallow	.63	[.35, .68]	[.53, .67]	[.58, .73]	[.48, .68]	[.60, .70]	[.57, .67]
	Deep	.81	[.51, .85]	[.76, .84]	[.74, .83]	[.69, .95]	[.80, .85]	[.75, .85]
	LLM	.60	[.50, .63]	[.46, .62]	[.55, .61]	[.51, .77]	[.59, .59]	[.58, .61]
Conclusion	Shallow	.55	[.44, .71]	[.50, .58]	[.46, .55]	[.46, .61]	[.54, .55]	[.52, .57]
	Deep	.70	[.64, .80]	[.59, .74]	[.63, .81]	[.64, .78]	[.64, .71]	[.64, .78]
	LLM	.68	[.63, .76]	[.68, .81]	[.63, .67]	[.58, .84]	[.65, .68]	[.61, .72]
Major Claim	Shallow	.68	[.62, .74]	[.66, .74]	[.49, .75]	[.42, .81]	[.66, .72]	[.62, .72]
	Deep	.88	[.78, .92]	[.87, .88]	[.80, .95]	[.81, .89]	[.87, .88]	[.84, .90]
	LLM	.75	[.68, .82]	[.66, .81]	[.63, .84]	[.71, .91]	[.71, .86]	[.66, .86]
Position	Shallow	.41	[.34, .46]	[.34, .53]	[.16, .49]	[.29, .56]	[.36, .50]	[.17, .58]
	Deep	.44	[.23, .56]	[.36, .73]	[.23, .61]	[.28, .46]	[.37, .59]	[.27, .54]
	LLM	.32	[.13, .37]	[.29, .54]	[.29, .47]	[.22, .60]	[.31, .33]	[.23, .37]
Warrant	Shallow	.43	[.32, .51]	[.39, .51]	[.38, .51]	[.38, .47]	[.39, .55]	[.37, .52]
	Deep	.44	[.27, .53]	[.38, .55]	[.36, .68]	[.36, .65]	[.41, .52]	[.25, .54]
	LLM	.00	[-.16, .09]	[-.02, .32]	[-.18, .02]	[-.04, .14]	[-.02, .07]	[-.13, .08]

Table 2: Kappa values for the individual classifiers evaluated either on all test essays or on essays from a certain subgroup. We report the minimal and maximal values among the subgroups for each demographic variable.

respective subgroup information being the independent variable for osa and osd. For csd, two models are fitted, one with both the subgroup and human score as independent variables and one using the human score only. We use the R^2 as a measure of model fairness for osa and osd and the difference in R^2 for csd. In our analysis we follow Williamson et al. who established that absolute values above 0.1 suggests unfairness or bias against certain groups.

Fairness should be considered in addition to mean accuracy because research on teacher judgments has shown that the qualities of judgments are almost uncorrelated, and teachers who are very good at judging the average class level can be very unfair to the high or low-performing students((Möller et al., 2022),(Urhahne and Wijnia, 2021)).

We used Cohen's kappa to account for chance agreement in evaluating our model. This is crucial when classifiers evaluate argumentative elements in essays. Percentage agreement alone may overstate accuracy by reflecting chance, misleading results. Kappa provides a more accurate measurement of agreement strength. This is crucial in educational settings, where precise feedback is necessary, as ignoring chance agreement could overestimate teacher judgments. By incorporating kappa, we aim for a more balanced evaluation of our classifier's performance and fairness across diverse student groups, improving feedback in educational technologies and reducing biases in teacher assessments.

5 Experimental Study

In the following, we discuss the results of our experimental studies. We compare the three classification model types (**Shallow**, **Deep**, and **LLM**) with respect to both fairness and kappa. In the first experiment, we trained on the complete dataset and evaluated the fairness for certain subgroups.

In a second experiment, we trained models on subsets of the training data that represent only a specific part of the whole population (in our case, the upper and lower quartiles of the cognitive ability values) and examined the fairness of such models.

5.1 Evaluation of Full Models on Fairness and Performance

Table 2 presents the performance of our trained models with regard to chance-corrected kappa values, providing insights into the agreement between model predictions and human annotators. The range values in brackets show variances across the different subgroups. We excluded the subgroup Aesthetic from the category Profile, as it had only 9 students and led to extreme outliers. Our study involved three machine learning models: Shallow (SVM), Deep (BERT), and LLM (gpt-4-turbo-preview, GPT). The prompts used for the LLM are displayed in the Appendix.

For the prediction of the Introduction the Deep model demonstrated the highest performance with an overall kappa of .81, indicating a strong agreement with human annotations. In contrast, the Shallow and LLM models performed worse, a trend that persists through all models. The order of the model

Label	Metric	Model	Grades	Gender	Profile	School	Language	KFT
Introduction	osa	Shallow	.008	.001	-.002	-.001	.000	-.001
		Deep	.011	-.002	-.003	.001	-.001	.003
		LLM	-.000	-.001	-.004	.000	-.001	-.002
	osd	Shallow	.005	.004	-.001	.010	-.001	.007
		Deep	.001	.005	.000	-.001	-.001	-.000
		LLM	.014	-.001	.004	-.000	-.000	.001
	csd	Shallow	.019	.026	.038	.013	.001	.012
		Deep	.009	.022	.037	.004	.001	.000
		LLM	.032	-.002	.014	-.007	-.000	.008
Conclusion	osa	Shallow	.014	-.001	-.003	-.001	.000	.000
		Deep	-.003	.000	.007	-.002	-.001	.001
		LLM	.005	-.001	-.004	.002	-.001	-.002
	osd	Shallow	.004	.001	.004	.002	-.000	-.002
		Deep	-.002	-.001	.001	.006	.002	-.001
		LLM	.001	-.000	.000	.003	-.001	-.002
	csd	Shallow	-.003	.005	.019	-.001	.005	.005
		Deep	-.000	-.004	-.024	.004	-.001	-.002
		LLM	.003	-.007	.014	.000	-.000	-.000
Major Claim	osa	Shallow	-.002	-.002	-.004	.006	-.001	-.001
		Deep	.001	-.002	-.001	-.002	-.001	-.000
		LLM	-.001	.004	-.001	.001	.003	.005
	osd	Shallow	.003	-.001	-.002	.001	-.001	.007
		Deep	-.001	-.001	-.003	.000	-.001	-.000
		LLM	.004	-.002	-.002	-.002	.001	-.002
	csd	Shallow	.002	-.010	.011	.007	-.001	.005
		Deep	-.002	.001	.004	-.001	-.001	.000
		LLM	.002	.005	.044	.008	.003	-.001
Position	osa	Shallow	-.003	-.001	.003	.015	.001	.008
		Deep	.003	-.000	-.003	.017	-.001	.005
		LLM	.005	-.001	.004	.001	.001	.008
	osd	Shallow	.005	-.002	.012	.007	.002	.003
		Deep	-.000	-.001	-.002	.007	.001	-.002
		LLM	.004	-.002	.006	.007	-.001	.002
	csd	Shallow	.000	.012	.057	.019	.001	.010
		Deep	.002	.019	.050	.018	-.000	.014
		LLM	.008	-.010	-.018	-.005	.002	.022
Warrant	osa	Shallow	.007	-.002	.003	-.001	.007	.006
		Deep	.007	.001	.018	.008	.004	.016
		LLM	.012	.004	.005	-.003	.003	.008
	osd	Shallow	.000	.004	.002	.009	-.001	.003
		Deep	-.001	.002	-.002	-.001	-.001	-.000
		LLM	-.001	-.002	-.004	.004	-.000	.006
	csd	Shallow	.010	.002	-.036	.003	.000	-.001
		Deep	-.001	.011	-.008	.005	-.001	-.002
		LLM	.008	.006	.086	.007	.005	.025

Table 3: Fairness evaluation metrics of all classifiers.

performance is also reflected in the results ordered by demographic data.

For the Conclusion, the Deep model similarly outperformed its counterparts again, followed closely by the LLM. The SVM stays behind. When evaluating Major Claim, all models display a noticeably enhanced performance, especially the Deep model, reaching a kappa value of .88 followed by the LLM (.75), and lastly the Shallow model .68.

For Position and Warrant, kappa values reveal a drop in performance across all models, with the Deep model followed closely by the SVM. The LLM model lags behind, for the Position annotation at a value around zero, showing challenges in capturing the nuanced expression of stances or viewpoints within texts. Those results seem to mirror also the inter-annotator agreements of the original annotation, in which the annotations for Introduction/Conclusion (content zone) and Major Claim had both an inter-annotator Krippendorffs alpha of .83, the Position annotation at .68, while all TAP values (e.g. warrant) showed very low agreements.

The analysis reveals the strengths and weaknesses inherent to each modeling approach. Deep learning models, particularly BERT, consistently demonstrated robust kappa scores, affirming their suitability for complex linguistic tasks. Depending on the task, the SVM varied between staying behind between 1 to 18 points from BERT. In contrast, the generative capabilities of LLM models, such as GPT, varied extremely in their performance, although never outperforming the Deep model. These findings underscore the importance of model selection based on the specific demands of the task at hand. It is entirely possible that different prompts would have led to different results. However, it would have to be examined whether the resources required (time to develop and test the appropriate prompt, cost of the queries, energy consumption of LLM models) justify this procedure.

Table 3 shows the fairness measures based on the models, trained on the whole dataset. As reported, values over .10 are potentially an issue of concern. None of the calculations on any model resulted in any value above .10.

5.2 Training Models on KFT Subgroups

As a second step, we estimated the effects it can have if certain subgroups are not adequately reflected in the training data. For this experiment, we considered specifically cognitive abilities represented by cognitive ability values. We divided the training data into four quartiles based on the cognitive ability values and trained models on data from the lowest and highest quartiles only. For a more balanced comparison to general data, we also sampled a comparable size of training data from all four quartiles in a stratified way, e.g. from each quartile we took a randomised sample of 25%. This subset is further referret to as mixed data. This experiment was not conducted for LLMs, as our zero-shot approach does not rely on training data.

Unsurprisingly, the performance of both the SVM and the BERT model deteriorated in comparison to models trained on the full training set (see Table 4).

In general, the deep model performed still better than the shallow one, except for the position model trained on the low quartile as well as the warrant models trained on the highest and lowest quartiles. There is no indication that any of the quartiles lead to a stronger model. Each category