Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education

Nils-Jonathan Schaller[1], **Yuning Ding**[2], **Andrea Horbach**[2,3], **Jennifer Meyer**[1], **Thorben Jansen**[1]
[1]Leibniz Institute for Science and Mathematics Education at the University of Kiel, Germany
[2]CATALPA, FernUniversität in Hagen, Germany
[3]Hildesheim University, Germany

Abstract

Pursuing educational equity, particularly in writing instruction, requires that all students receive fair (i.e., accurate and unbiased) assessment and feedback on their texts. Automated Essay Scoring (AES) algorithms have so far focused on optimizing the mean accuracy of their scores and paid less attention to fair scores for all subgroups, although research shows that students receive unfair scores on their essays in relation to demographic variables, which in turn are related to their writing competence. We add to the literature arguing that AES should also optimize for fairness by presenting insights on the fairness of scoring algorithms on a corpus of learner texts in the German language and introduce the novelty of examining fairness on psychological and demographic differences in addition to demographic differences. We compare shallow learning, deep learning, and large language models with full and skewed subsets of training data to investigate what is needed for fair scoring. The results show that training on a skewed subset of higher and lower cognitive ability students shows no bias but very low accuracy for students outside the training set. Our results highlight the need for specific training data on all relevant user groups, not only for demographic background variables but also for cognitive abilities as psychological student characteristics.

1 Introduction

Educational equity is seen as a foundation for learning with technology (Warschauer et al., 2004), because all students need effective instruction. One of the most effective instructional practices is feedback (Hattie and Timperley, 2007), which can support students in acquiring complex skills like writing (Graham et al., 2015). Automated essay scoring (AES) can be used to provide students with feedback on their writing at scale (Fleckenstein et al., 2023).

The foundation of equity in automated feedback systems is the fairness of the algorithm ((Holstein and Doroudi, 2021), (Pedró et al., 2019)), i.e., the absence of any prejudice or favoritism toward groups of students based on their inherent or acquired characteristics, including their background and their psychological variables((Mehrabi et al., 2019),(Government Equalities Office, 2013)). Algorithmic fairness is widely discussed in various educational contexts from normative (Blodgett et al., 2020; European Commission, Directorate-General for Education, Youth, Sport and Culture, 2022), societal (Baker and Hawn, 2022; Kizilcec and Lee, 2020), or methodological (Mitchell et al., 2021) perspectives, but literature reviews have shown that it is rarely investigated empirically (Li et al., 2023). Specifically in the AES context, only six empirical studies have examined algorithmic fairness, examining differences in algorithmic accuracy and biases for students with different gender, race, and language backgrounds in English-language corpora (Arthur et al., 2021; Baffour et al., 2023; Bridgeman et al., 2009; Litman et al., 2021; Kwako et al., 2022; Yancey et al., 2023). This means that while AES is widely used in education in many countries (Fleckenstein et al., 2023) including non-English speaking countries, it is unclear whether the algorithms used are fair to all groups of students confronted with the results or whether they might disfavor some student gropus. Compounding the problem, the few existing studies have shown that, depending on the algorithms used, students' essays were not scored fairly and disfavored groups related to race/ethnicity, economic status, and English Language Learner status (e.g., Baffour et al. (2023); Litman et al. (2021); Yang et al. (2024)).

So far, previous studies only analyzed fairness in relation to students' demographic variables in corpora with students' essays in English: Extending this research to a corpus on argumentation essays in the German language, we address three main re-

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 210–221
June 20, 2024 ©2024 Association for Computational Linguistics

search questions: (1) How fair are AES algorithms for students with different levels of cognitive abilities as psychological characteristics strongly related to writing competence? (Zhang and Zhang, 2023). Addressing this question is linked to the wider equity issue of whether AES systems are likely to widen or narrow the gap between high and low-performing students. (2) How fair are AES algorithms in languages other than English? The question is especially important when automated scoring is based on large language models, mostly trained on English text data. (3) How is the distribution of student characteristics in the training data impacting the mean accuracy and fairness of the prediction?

By answering these questions, our paper makes the following contributions: First, we provide a set of baseline models, including shallow learning, deep learning, and generative large language models (LLM), for the newly released DARIUS corpus, thus enriching the automatic scoring landscape with models for a large German argumentative writing corpus.

Second, we conduct fairness evaluations on our results indicating that none of the models trained on the entirety of training data shows unfair behavior towards specific subgroups.

Finally, to assess the role of the distribution of the training data on algorithmic fairness, we train shallow and deep models with subsets of data from students of low and high cognitive ability, as well as a mixed subset based on low, medium and high cognitive ability, and show that the models are unfair to the groups not included in the training set.

We make all of our code publicly available.[1]

2 Related Work: Fairness in AES Algorithm

According to a literature review by Li et al. (2023), there have been 49 peer-reviewed empirical studies focused on fairness and predictive bias in education since 2010, highlighting the growing academic interest in these issues.

The studies included multiple fairness measures, including the accuracy for the included groups and the mean differences between predicted and annotated scores for each score (e.g., (Litman et al., 2021)). Most of these studies were conducted in contexts other than AES, such as predicting students' course performance or their likelihood of dropping out of a course. To our knowledge, there are only two papers that diagnosed the predictive bias displayed by AES models(Litman et al., 2021; Arthurs and Alvero, 2020), even though the importance of this task has been pointed out as early as in 2012 (Williamson et al., 2012). Litman et al. (2021) evaluated the fairness of shallow and deep learning AES algorithms for essays from the upper elementary level in the English language using three measures: Overall Score Accuracy (OSA), Overall Score Difference (OSD), and Conditional Score Difference (CSD). They found that shallow and deep AES algorithms showed systematically overly positive and negative scoring depending on students' gender, race, and socioeconomic status. Arthurs and Alvero (2020) showed that a shallow learning AES system for college admissions essays based on word vectors favored high-income students over low-income students (see also (Bridgeman et al., 2009) for similar results for essays from the Test of English as a foreign language). Additionally, the authors trained models on only essays from the highest quartile of students in terms of performance, showing that these models are not suitable for students from the other quartiles. Yang et al. (2024) further emphasized that the fairness of AES systems is compromised if such models are used on students or tasks for which they have not been trained.

In addition to the studies included in the literature review, recent studies added an investigation of fairness in Large Language Models scoring essays from a high school context Baffour et al. (2023) in the PERSUADE 2.0 corpus (Crossley et al., 2022). The authors compared the winning entries of the Kaggle Feedback Prize competition.[2] They show differences in the model's accuracy based on demographic factors such as student race/ethnicity, and economic disadvantage. Similar fairness issues based on students' demographic variables were shown for large language models in essays in the English language written by first (Kwako et al., 2023) and second language students (Yancey et al., 2023).

In summary, previous studies on fairness in AES have used shallow learning models, deep learning models, and LLMs and compared whether the accuracy of judgments and systematic over/underrating can be explained by students' demographic vari-

[1] https://github.com/darius-ipn/fairness_AES

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

ables. The results showed some fairness problems, which were exacerbated in the studies where the AES was additionally trained only on a homogenous group of students.

3 Data

The DARIUS corpus is a collection of 4,589 annotated argumentative texts written by 1,839 students from German high schools, spread across 114 classes in 33 different schools(Schaller et al., 2024). Essays that were off-topic, shorter than two sentences, empty, or contained names or other data relevant to data protection were removed beforehand. The final dataset consists of essays from two writing assignments focused on socio-scientific issues on the topics *energy* and *automotive*, containing 2,307 and 2,282 essays respectively. Students wrote a draft and revision on one task, followed by an essay on the other task, resulting in up to 3 essays per student. An example text is listed in the Appendix 7. Students also provided demographic data voluntarily, a selection of which is listed in Table 1.

The dataset has been extensively annotated with information about argumentative structure on different levels of granularity. In the present study, we focus specifically on a subset of these annotations, namely *content zone*, *major claim*, *position* and *warrant*. Out of the nine original annotation categories, we selected those as they reflect different parts of an argumentative text, e.g. structure and content, and are annotated on different granularity levels (token level to whole texts). We used the demographic data to measure fairness with respect to gender, profile, school, cognitive ability (KFT), and languages, which are further explained after providing more details on the annotations in Section 3.1. A more extensive description can be found in the original paper (Schaller et al., 2024).

3.1 Annotations

Content zone: This annotation category breaks down the essays into their basic parts: the introduction, the body, and the conclusion. Each section can be as short as one sentence or span several sentences.

Major claim annotation: Central to the argumentative essence of the essays, the Major Claim annotation identifies the pivotal stance taken by the author on the discussed issue. In contrast to similar annotation efforts (Stab and Gurevych, 2014), we also include claims written not only in the opening paragraphs but also within the conclusion, offering a comprehensive view of the argumentative intent. Such claims form the basis for the author's further arguments and the direction of their reasoning.

Position annotation: This annotation extracts the essay's directional stance regarding the thematic issues presented in the writing tasks — whether the argumentation aligns with, diverges from, or remains ambiguous towards the positions debated within the tasks. This annotation is important for understanding the diversity of viewpoints and the critical engagement of students with the socio-scientific topics at hand.

Warrant annotation: A warrant is one out of five argumentative elements annotated in the dataset as part of the Toulmin's Argumentation Pattern (TAP) annotations, following the definitions by Riemeier et al. (2012). TAP describes a structural framework for constructing logical and compelling arguments by including a claim, providing supporting evidence (data), explaining the connection between the claim and data (warrant), and addressing counterarguments (rebuttal). For this study, we focus exemplarily on warrants because the use of warrants indicates already a higher argumentation skill(Osborne et al., 2016). TAP elements are not marked on the sentence level but on the token level, as a TAP sequence can cover a wide range from subordinate clauses to entire paragraphs.

3.2 Demographic and Psychological Data

We consider the following demographic variables: **Grade** Grade indicates which grade level the student is in. The dataset was obtained for students between Grade 9 and Grade 12.

Gender The students could indicate their gender. Options were female, male, and diverse.

School The German school system differentiates between different forms of high school.

- Gemeinschaftsschule: non-academic track
- Gymnasium: academic track
- Berufsschule: vocational training

Profile The German school system allows students to choose a profile. The Natural Sciences profile, for example, has a focus on math and science, while the Social Sciences profile can have a focus on politics or ethics.

Languages The students could indicate the language that they speak at home.

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.

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

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

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

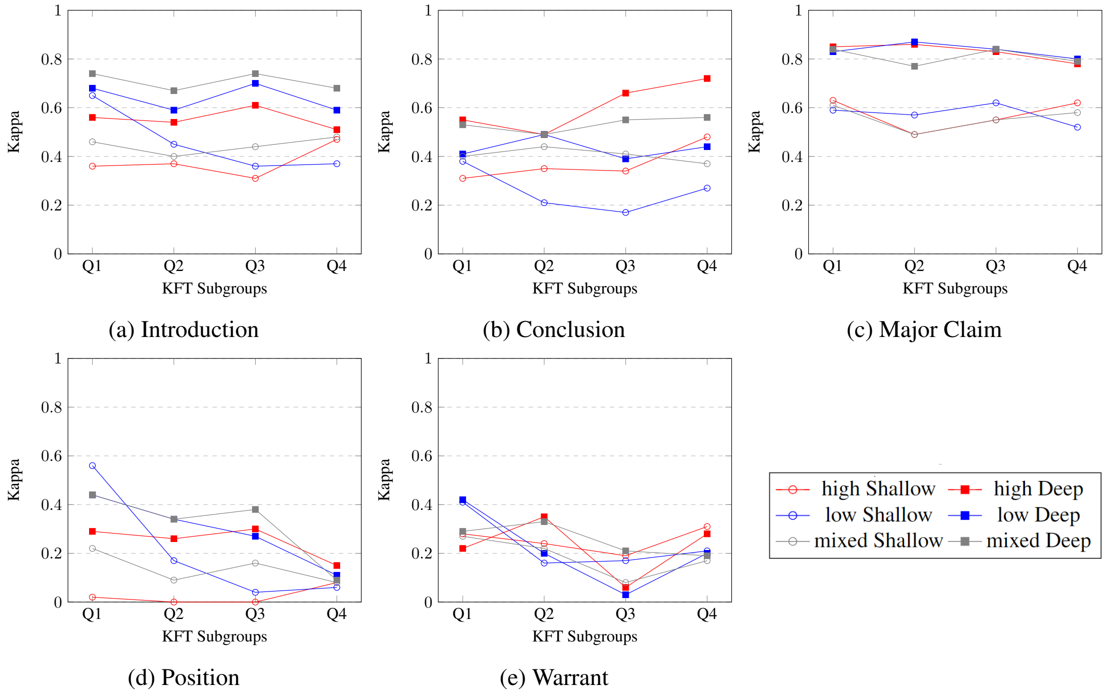
As a second step, we estimated the effects it can have if certain subgroups are not adequately reflected in the training data. For this experiment, we considered specifically cognitive abilities represented by cognitive ability values. We divided the training data into four quartiles based on the cognitive 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

Label	KFT	Model	All	Grades	Gender	Profile	School	Languages
Introduction	high	Shallow	.38	[-.04, .44]	[.33, .62]	[.29, .39]	[.18, .48]	[.35, .45]
		Deep	.56	[.30, .62]	[.29, .59]	[.45, .56]	[.25, .57]	[.53, .61]
	low	Shallow	.47	[.26, .48]	[.30, .43]	[.30, .65]	[.41, .57]	[.40, .64]
		Deep	.65	[.59, .67]	[.63, .68]	[.60, .71]	[.62, .70]	[.64, .64]
	mixed	Shallow	.46	[.06, .51]	[.39, .61]	[.40, .47]	[.17, .55]	[.41, .57]
		Deep	.71	[.65, .73]	[.68, .71]	[.70, .76]	[.70, .73]	[.70, .75]
Conclusion	high	Shallow	.39	[.21, .48]	[.37, .53]	[.29, .47]	[.21, .52]	[.27, .40]
		Deep	.62	[.49, .66]	[.56, .65]	[.53, .72]	[.52, .77]	[.58, .62]
	low	Shallow	.25	[.19, .27]	[.17, .28]	[.21, .23]	[.09, .29]	[.20, .25]
		Deep	.44	[.16, .51]	[.40, .43]	[.29, .47]	[.34, .62]	[.41, .54]
	mixed	Shallow	.42	[.32, .55]	[.41, .56]	[.34, .44]	[.35, .45]	[.34, .42]
		Deep	.54	[.43, .57]	[.49, .69]	[.50, .63]	[.45, .62]	[.54, .54]
Major Claim	high	Shallow	.57	[.47, .63]	[.50, .62]	[.36, .58]	[.35, .57]	[.55, .61]
		Deep	.83	[.67, .87]	[.81, .88]	[.79, .90]	[.78, .92]	[.82, .85]
	low	Shallow	.58	[.46, .63]	[.52, .62]	[.37, .67]	[.35, .66]	[.57, .61]
		Deep	.84	[.77, .89]	[.80, .86]	[.76, .95]	[.70, .85]	[.82, .87]
	mixed	Shallow	.56	[.49, .62]	[.52, .56]	[.31, .70]	[.35, .58]	[.52, .67]
		Deep	.81	[.58, .86]	[.70, .82]	[.73, .89]	[.61, .82]	[.79, .87]
Position	high	Shallow	.02	[.00, .05]	[.00, .03]	[.00, .00]	[.00, .04]	[.00, .03]
		Deep	.29	[-.05, .43]	[.23, .49]	[-.04, .43]	[.17, .43]	[.27, .30]
	low	Shallow	.37	[.34, .48]	[.28, .69]	[.29, .41]	[.19, .69]	[.28, .52]
		Deep	.34	[-.07, .40]	[.28, .71]	[.23, .47]	[.08, .61]	[.29, .44]
	mixed	Shallow	.16	[.00, .18]	[.00, .15]	[.06, .15]	[.00, .37]	[.14, .18]
		Deep	.37	[-.03, .43]	[.33, .53]	[.24, .43]	[.29, .43]	[.33, .43]
Warrant	high	Shallow	.26	[.10, .32]	[.21, .29]	[.23, .29]	[.05, .27]	[.23, .36]
		Deep	.23	[.13, .30]	[.18, .31]	[.21, .34]	[.16, .37]	[.21, .29]
	low	Shallow	.23	[.19, .24]	[.19, .30]	[.20, .28]	[.14, .35]	[.19, .37]
		Deep	.20	[.03, .26]	[.16, .34]	[.19, .41]	[.12, .61]	[.16, .34]
	mixed	Shallow	.17	[.13, .22]	[.16, .28]	[.12, .31]	[.05, .41]	[.16, .22]
		Deep	.25	[.18, .30]	[.20, .39]	[.22, .28]	[.22, .49]	[.24, .29]

Table 4: Kappa values of KFT classifiers and all subtypes.



(a) Introduction

(b) Conclusion

(c) Major Claim

(d) Position

(e) Warrant

Figure 1: Kappa values of KFT classifiers on different KFT subgroups. Q = Quartile.

Label	Metric	KFT	Model	Grades	Gender	Profile	School	Language	KFT
Introduction	osa	high	Shallow	.011	.005	-.004	.003	-.001	.001
			Deep	.001	.003	-.002	.000	-.001	.001
		low	Shallow	.001	.001	.007	.000	.008	.007
			Deep	-.003	-.001	-.003	-.002	-.001	.003
		mixed	Shallow	.008	-.001	-.004	.003	.001	.001
			Deep	-.003	-.000	-.004	-.003	-.001	.000
	osd	high	Shallow	.003	.003	-.004	.000	.001	-.000
			Deep	.001	-.001	.005	.000	-.000	.001
		low	Shallow	.002	.000	.003	.002	-.001	.006
			Deep	-.002	.007	.004	.002	-.001	.002
		mixed	Shallow	.010	-.002	.000	.009	-.001	.000
			Deep	-.001	.000	.002	.006	-.001	.004
	csd	high	Shallow	.012	.017	.093	.016	.000	-.001
			Deep	.014	.007	.074	.007	.007	.006
		low	Shallow	.018	.025	.053	.020	.005	.013
			Deep	.011	.008	.034	-.005	.002	.006
		mixed	Shallow	.022	.017	.065	.022	.002	.009
			Deep	.009	.011	.015	.004	-.000	.010
Conclusion	osa	high	Shallow	.011	.001	.003	.004	-.001	.002
			Deep	.000	-.000	.002	.001	-.001	.004
		low	Shallow	.010	-.000	-.004	.001	.002	.016
			Deep	.006	-.002	.000	-.001	.006	-.000
		mixed	Shallow	.011	.001	-.003	-.002	-.001	.001
		mixed	Deep	-.001	.003	-.001	-.001	-.001	-.003
	osd	high	Shallow	.016	-.002	.005	.004	-.001	-.001
			Deep	-.004	.002	-.004	-.003	-.000	-.001
		low	Shallow	.004	-.002	-.002	.000	.003	.012
			Deep	.005	-.002	.001	-.000	-.001	-.003
		mixed	Shallow	.003	-.002	-.000	.001	-.001	-.003
			Deep	-.001	-.001	.003	-.002	-.001	-.002
	csd	high	Shallow	.010	-.009	-.025	-.007	.006	.010
			Deep	.004	.003	-.007	-.003	-.001	.004
		low	Shallow	.001	.006	-.033	.006	-.000	-.001
			Deep	.004	.012	.042	.008	.001	.002
		mixed	Shallow	.001	-.009	-.003	-.011	.004	.003
			Deep	.004	.006	.034	.004	.001	.007
Major Claim	osa	high	Shallow	-.001	.004	-.004	.008	-.001	.000
			Deep	.001	-.002	-.003	.004	-.001	-.003
		low	Shallow	-.002	.003	-.001	.003	-.001	-.003
			Deep	.001	-.000	.002	-.001	-.000	-.002
		mixed	Shallow	.000	-.001	-.001	.018	.000	-.001
			Deep	.006	.001	.003	.003	.001	-.002
	osd	high	Shallow	-.001	.000	-.002	-.003	-.000	.005
			Deep	-.002	-.000	-.004	-.003	-.000	-.002
		low	Shallow	.004	.002	.002	.000	-.001	.000
			Deep	.003	-.000	-.004	-.002	.000	-.000
		mixed	Shallow	.006	-.002	-.002	-.003	-.001	.003
			Deep	.002	-.001	-.001	-.001	-.001	-.001
	csd	high	Shallow	-.002	-.004	.032	.014	-.001	.005
			Deep	-.002	.006	-.010	.005	.001	-.002
		low	Shallow	.002	.002	.043	.014	-.001	-.000
			Deep	.002	-.001	-.003	-.005	-.000	-.000
		mixed	Shallow	.005	.000	.020	.021	-.000	.004
			Deep	.002	.000	.012	.004	-.001	-.001
Position	osa	high	Shallow	.003	.002	.020	.011	.014	.036
			Deep	-.002	-.002	-.002	.012	.007	.024
		low	Shallow	.002	-.000	.007	.016	.000	.015
			Deep	-.001	-.001	-.000	.003	.001	.005
		mixed	Shallow	.001	.003	.016	.008	.009	.026
			Deep	-.001	-.002	.020	.009	.002	.011
	osd	high	Shallow	.003	.002	.020	.011	.014	.036
			Deep	.000	-.000	-.001	.006	.001	.003
		low	Shallow	.000	-.002	.005	.006	-.001	.006
			Deep	-.003	-.002	-.001	.005	.000	.003
		mixed	Shallow	.002	.003	.017	.006	.010	.024
			Deep	-.002	-.002	.005	.001	.003	.002
	csd	high	Shallow	-.000	-.003	.015	-.003	.000	-.000
			Deep	.003	-.013	.096	-.014	.001	.004
		low	Shallow	-.001	.030	.027	.039	.005	.013
			Deep	.001	.008	.017	.016	.001	.003
		mixed	Shallow	-.001	.019	.041	.025	-.000	.001
			Deep	.002	.005	.059	.005	-.000	.004
Warrant	osa	high	Shallow	.010	-.001	-.000	-.002	.006	.004
			Deep	.003	-.000	.003	.007	.004	.015
		low	Shallow	.014	-.002	.001	-.000	.008	.010
			Deep	.011	-.001	.012	.013	.008	.023
		mixed	Shallow	.019	-.002	.013	.007	.004	.015
			Deep	.007	.001	.002	.001	.003	.009
	osd	high	Shallow	.005	-.002	-.002	.004	-.001	.001
			Deep	.000	-.000	-.003	-.002	-.001	-.001
		low	Shallow	.003	-.001	.016	.011	.001	.013
			Deep	.005	-.002	.002	.001	.000	.001
		mixed	Shallow	.012	-.002	.009	.011	-.000	.007
			Deep	.005	-.002	-.002	-.001	-.001	.002
	csd	high	Shallow	.002	-.003	-.047	-.002	.000	.003
			Deep	.009	.007	-.020	.002	.002	-.000
		low	Shallow	.003	-.008	-.042	-.001	-.000	.003
			Deep	-.000	-.005	-.035	-.002	-.001	-.001
		mixed	Shallow	.001	-.017	-.045	-.007	.000	-.001
			Deep	.000	-.011	-.014	-.008	.002	.001

Table 5: Fairness evaluation metrics of KFT classifiers and all subtypes.

(low, high, and mixed) can perform best in different tasks, e.g. *mixed deep* in Introduction, *high deep* in Conclusion, or *low shallow*/*mixed deep* in Position. In terms of fairness, we still found no values above 0.1 (see Table 5).

When examining Figure 1 we can see that models differed in their performance when tested on different subgroups. For the Introduction, a shallow model trained on the dataset of the students with the highest KFT quartile (*high shallow*) was performing better on the subgroup it was trained on (e.g. Quartile 4) than on the other subgroups and the other way around (low KFT model performed better on the subset with low KFT, e.g. Quartile 1.). The mixed models had the lowest variance in performance.

There are exceptions in which the model performed better on a different subgroup than the one it was trained on, e.g., in (d) Position, all models except high shallow lost performance on Quartile 4. Furthermore, all combinations of algorithm and training data did have a comparable stable performance on (c) Major Claim.

In general, using training data from only one student group seemed to introduce a bias, disadvantaging other student groups. This finding underlines the need to include training data from a diverse range of students to ensure fairness and avoid skewed outcomes.

6 Conclusion and Future Work

In our work, we provide three basic models (shallow learning models, deep learning models, and LLM) trained on the annotations of the DARIUS corpus of learner texts in German. These models are ready to use in schools, for example, to create a feedback tool for training argumentative skills. Evaluation of model fairness showed that all models produced fair scores for all students, considering demographic and psychological differences among students. In a second experiment, we trained our models on subgroups of students, based on either low, high, or mixed cognitive abilities, to investigate the extent to which skewed training data leads to unfair AES system scores. Our results showed lower performance for students who were not in the training data, emphasizing the importance of including samples of the full range of users in the training data for AES, not only for demographic background variables but also for psychological aspects such as cognitive abilites. Failure to do so risks reducing the predictive accuracy of the algorithm for those who are not adequately represented. To mitigate the risk of students receiving unfair scores based on their demographic and psychological variables, we advocate that future AES systems incorporate the goal of fairness in addition to accuracy into their training data collection and algorithm optimization function, going beyond the current state of retrospective analysis of model fairness.

7 Limitations

This study encounters several limitations that have to be mentioned. One constraint is the small size of certain subgroups within the corpus, as seen in Table 1, e.g., students with specific family languages, profiles like Linguistics or Aesthetics. The underrepresentation of those subgroups poses a challenge in drawing robust conclusions for these particular groups, potentially impacting the reliability and applicability of our outcomes to these populations.

Additionally, the comparatively homogenous population in the state of Schleswig-Holstein in northern Germany, restricts the generalizability of our findings. The demographic profile of Schleswig-Holstein may not reflect the diversity found in other regions or countries, potentially narrowing our study's insights.

In conclusion, while our study provides insights into fairness in the subgroups of the DARIUS Corpus, these limitations underscore the necessity for a cautious interpretation of our findings and suggest areas for future research efforts to build upon and address these constraints.

8 Acknowledgements

This work was supported by the Deutsche Telekom Stiftung and partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

References

Philip Arthur, Dongwon Ryu, and Gholamreza Haffari. 2021. Multilingual simultaneous neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4758–4766, Online. Association for Computational Linguistics.

Noah Arthurs and AJ Alvero. 2020. Whose truth is the "ground truth"? college admissions essays and bias in word vector evaluation methods. *International Educational Data Mining Society*.

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246.

Ryan S. Baker and Aaron. Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32:1052–1092.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2009. Considering fairness and validity in evaluating automated scoring. In *Annual Meeting of the National Council on Measurement in Education*, San Diego, CA.

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.

European Commission, Directorate-General for Education, Youth, Sport and Culture. 2022. *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*. Publications Office of the European Union.

Johanna Fleckenstein, Lucas W. Liebenow, and Jennifer Meyer. 2023. Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6.

Government Equalities Office. 2013. Equality Act 2010: guidance. `https://www.gov.uk/guidance/equality-act-2010-guidance`. Accessed: 2023-09-21.

Steve Graham, Michael Hebert, and Karen Harris. 2015. Formative assessment and writing: A meta-analysis. *The Elementary School Journal*.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*.

Kurt Heller and Christoph Perleth. 2000. *Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT 4-12+ R)*.

Kenneth Holstein and Shayan Doroudi. 2021. Equity and artificial intelligence in education: Will "aied" amplify or alleviate inequities in education? *CoRR*, abs/2104.12920.

René F. Kizilcec and Hansol Lee. 2020. Algorithmic fairness in education. *CoRR*, abs/2007.05443.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. Using item response theory to measure gender and racial bias of a BERT-based automated English speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 1–7, Seattle, Washington. Association for Computational Linguistics.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang, and Li Cai. 2023. Does bert exacerbate gender or l1 biases in automated english speaking assessment? In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 668–681.

Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. 2023. Moral machines or tyranny of the majority? a systematic review on predictive bias in education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 499–508.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Matsumura, and Elaine L. Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In *International Conference on Artificial Intelligence in Education*, pages 255–267, Cham. Springer International Publishing.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Nitin Madnani and Anastassia Loukina. 2016. RSMTool: A collection of tools for building and evaluating automated scoring models. *Journal of Open Source Software*, 1(3).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and its Application*.

Jens Möller, Thorben Jansen, Johanna Fleckenstein, Nils Machts, Jennifer Meyer, and Raja Reble. 2022. Judgment accuracy of german student texts: Do teacher experience and content knowledge matter? *Teaching and Teacher Education*, 119:103879.

OpenAI. 2024. Gpt-4 technical report.

Jonathan Osborne, Bryan Henderson, Anna Macpherson, Evan Szu, Andrew Wild, and Shi-Ying Yao. 2016. The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Francesc Pedró, Miguel Subosa, Axel Rivas, and Paula Valverde. 2019. Artificial intelligence in education: challenges and opportunities for sustainable development. Working papers on education policy 7, UNESCO, France. Includes bibliography.

Tanja Riemeier, Claudia Aufschnaiter, Jan Fleischhauer, and Christian Rogge. 2012. Argumentationen von schülern prozessbasiert analysieren: Ansatz, vorgehen, befunde und implikationen. *Zeitschrift für Didaktik der Naturwissenschaften*, 18:141–180.

Nils-Jonathan Schaller, Andrea Horbach, Lars Höft, Yuning Ding, Jan L Bahr, Jennifer Meyer, and Thorben Jansen. 2024. Darius: A comprehensive learner corpus for argument mining in german-language essays. OSF Preprints. Accepted for LREC-COLING 2024.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Detlef Urhahne and Lisette Wijnia. 2021. A review on the accuracy of teacher judgments. *Educational Research Review*, 32.

Mark Warschauer, Michele Knobel, and Leeann Stone. 2004. Technology and equity in schooling: Deconstructing the digital divide. *Educational Policy*.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31.

Kevin P. Yancey, Geoffrey T. LaFlair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 576–584. Association for Computational Linguistics.

Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2024. Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jianhua Zhang and Lawrence Jun Zhang. 2023. Examining the relationship between english as a foreign language learners' cognitive abilities and l2 grit in predicting their writing performance. *Learning and Instruction*, 88.

A GPT prompts used in our experiments

Item	**Description**
Conclusion	Does this text have a concluding section, a summary? Answer with 1 for Yes or 0 for No.
Introduction	Does this text have an introduction? Answer with 1 for Yes or 0 for No.
Main Thesis	Is this text a main thesis, meaning a sentence in a text that takes a clear position? Answer with 1 for Yes or 0 for No.
Position	Does this text discuss all three positions of the task? Either cars that are powered by hydrogen, electricity, or e-fuels, or other task that involves hydroelectric power plants, solar power plants, and wind farms. If all three options are discussed, answer with 1, if not then 0.
Warrant	Do the arguments in the text have an explanation, meaning a more detailed explanation of the argument? If yes answer with 1, if not then 0.

Table 6: GPT prompts

B DARIUS corpus example

Deutsch	Englisch
In Norddeutschland wird die Frage gestellt welche klimaneutrale Energiegewinnung gebaut werden soll, um eine Klimaneutralität zu erreichen. Zur Frage kommen Windparks, Solar und Wasserkraftanlagen. Ich finde, dass der Bau von Windparks gefördert werden soll. Mit 45% Wirkungsgrad sind diese schwächer als Wasserkraftanlagen und stärker als Solarparks. Obwohl der Wirkungsgrad mit 45% geringer ist als bei Wasserkraftanlagen, liefert ein Windpark mit 40 GWh pro Jahr mehr Strom als Solarpark und Wasserkraftanlage. Ebenfalls ist der Preis relativ zum Jahresertrag günstig mit 14 Millionen als Solarpark und Wasserkraftanlage. Ebenfalls muss man in Betracht ziehen, dass der Windpark weniger CO2 ausstoßt. Solarpark und Wasserkraftanlage stoßen 35000t und 12000t CO2 und der Windpark nur 8,800t. Jedoch muss man sagen, dass der Windpark nur eine Lebensdauer von 20 Jahren hat. Währenddessen halten Solarparks 30 Jahre und Wasserkraftanlage 80 Jahre. Auf der Ebene der Lokalemissionen besitz der Windpark die meisten Emission mit Hör-, Infraschall und Schattenwurft. Die Wasserkraftanlage wirft keinen Schattenwurf, aber hat trotzdem Hör- und Infraschall. Der Solarpark hat keinen Emissionen jeglicher Art. Zum Schluss komme ich, dass man Windparks fördern sollte, da die Vorteile die Nachteile überwiegen. Sie bieten günstig Strom und verursachen wenig Treibhausgasemissionen, aber man muss anmerken, dass ein Windpark keine hohe Lebensdauer hat, sodass diese öfters erneuert werden müssen, und dass Anwohner und Tiere von diesem belästigt werden können.	In northern Germany, the question is being asked as to which climate-neutral energy generation should be built in order to achieve climate neutrality. The options are wind farms, solar and hydropower plants. I think that the construction of wind farms should be promoted. At 45% efficiency, they are less efficient than hydropower plants and more efficient than solar parks. Although the efficiency of 45% is lower than that of hydropower plants, a wind farm with 40 GWh per year supplies more electricity than solar farms and hydropower plants. The price relative to the annual yield is also lower at 14 million than solar parks and hydroelectric power plants. It must also be taken into account that the wind farm emits less CO2. The solar park and hydropower plant emit 35,000 tons and 12,000 tons of CO2 respectively, while the wind park emits only 8,800 tons. However, it must be said that the wind farm only has a lifespan of 20 years. In contrast, solar parks last 30 years and hydroelectric power plants 80 years. On the level of local emissions, the wind farm has the most emissions with acoustic, infrasound and shadow flicker. The hydropower plant does not cast any shadows, but still has audible and infrasound emissions. The solar park has no emissions of any kind. In conclusion, I believe that wind farms should be promoted because the advantages outweigh the disadvantages. They provide cheap electricity and cause little greenhouse gas emissions, but it should be noted that a wind farm does not have a long lifespan, so they have to be renewed frequently, and that residents and animals can be disturbed by them.

Table 7: Example essay in the DARIUS Corpus, translated via DeepL[4]