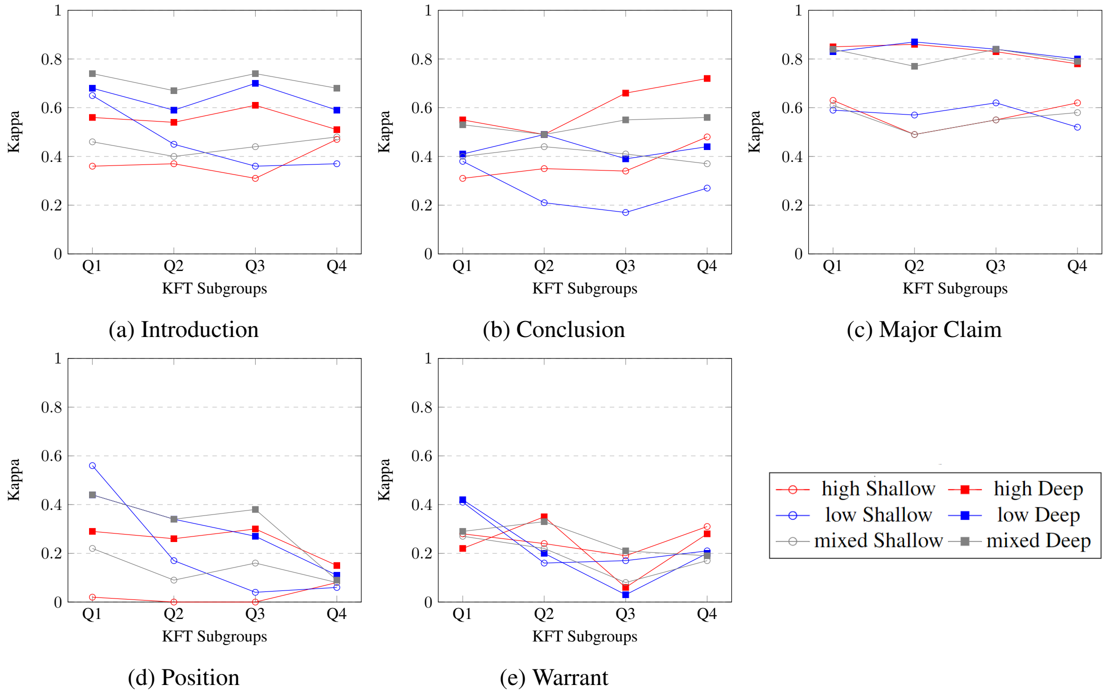
Label	KFT	Model	All	Grades	Gender	Profile	School	Languages
Introduction	high	Shallow	.38	[-.04, .44]	[.33, .62]	[.29, .39]	[.18, .48]	[.35, .45]
		Deep	.56	[.30, .62]	[.29, .59]	[.45, .56]	[.25, .57]	[.53, .61]
	low	Shallow	.47	[.26, .48]	[.30, .43]	[.30, .65]	[.41, .57]	[.40, .64]
		Deep	.65	[.59, .67]	[.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.