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]