

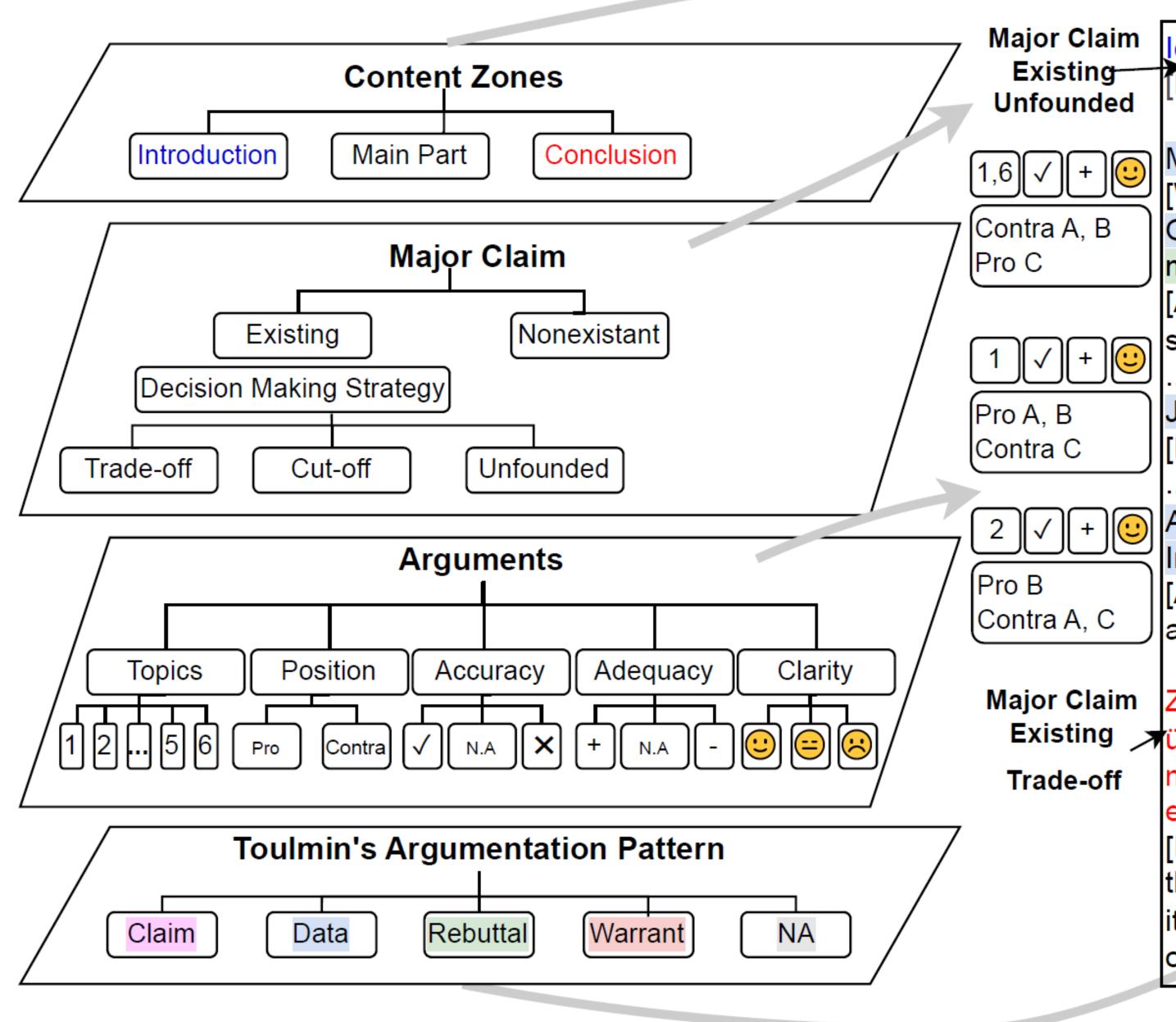
# Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education



DARIUS Digital Argumentation Instruction for Science

## Nils-Jonathan Schaller<sup>1</sup>, Yuning Ding<sup>2</sup>, Andrea Horbach<sup>2,3</sup>, Jennifer Meyer<sup>1</sup>, Thorben Jansen<sup>1</sup>





Ich finde, dass der Bau von Windparks gefördert werden soll. think that the construction of wind farms should be encouraged.]

Mit 45% Wirkungsgrad sind diese schwächer als Wasserkraftanlagen und stärker als Solarparks. [With 45% efficiency, these are weaker than hydroelectric plants and stronger than solar parks.] Obwohl der Wirkungsgrad mit 45% geringer ist als bei Wasserkraftanlagen, liefert ein Windpark mit 40 GWh pro Jahr mehr Strom als Solarpark und Wasserkraftanlage.

[Although the efficiency of 45% is lower than that of hydroelectric power plants, a wind farm supplies 40 GWh per year more electricity than solar farms and hydropower plants.]

Jedoch muss man sagen, dass der Windpark nur eine Lebensdauer von 20 Jahren hat. [However, it must be said that the wind farm only has a lifespan of 20 years.]

Auf der Ebene der Lokalemissionen besitz der Windpark die meisten Emission mit Hör-, Infraschall und Schattenwurft. Der Solarpark hat keinen Emissionen jeglicher Art. [At the level of local emissions, the wind farm has the most emissions with auditory, infrasound and shadows. The solar park has no emissions of any kind.]

Zum Schluss komme ich, dass man Windparks fördern sollte, da die Vorteile die Nachteile Existing wüberwiegen. 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 the end, I conclude that wind farms should be promoted because the advantages outweigh

the disadvantages. They provide cheap electricity and cause few greenhouse gas emissions, but it must be noted that a wind farm does not have a long lifespan, so they must be replaced more often, and residents and animals can be bothered by it.]

### Research Questions and Method

- How fair are AES algorithms for students with different levels of cognitive abilities as psychological characteristics strongly relate to writing competence?
- How fair are AES algorithms in languages other than English?
- How is the distribution of student characteristics in the training data impacting the mean accuracy and fairness of the prediction?

**Annotation and Metadata Selection** 

**Position** 

Contra;

Unclear

Argument stance.

- 1. Trained models on the complete dataset for each annotation.
- 2. Trained models on subsets of the training data based on cognitive abilities:
- Lower quartile of KFT
- Upper quartile of KFT

**Content Zone** 

Sequences of

9TH

10TH

introduction; main

part; conclusion.

- Mixed (quartile of each quartile) KFT
- 3. Evaluated the fairness of all models on a number of subgroups.

**Major Claim** 

Sequence of

12TH

13TH

topic.

**GRADE DISTRIBUTION** 

11TH

author's stance on a

## Fairness Measures and Corpus

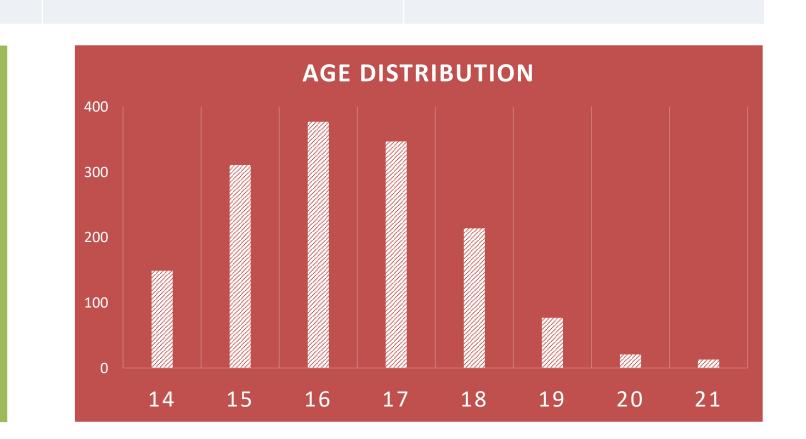
- overall score accuracy (osa): Model score vs human score across groups. (S H)<sup>2</sup>
- overall score difference (osd): How and whether some groups tend to score consistently higher or lower due to the model's predictions. (S-H)
- conditional score difference (csd): measures whether an AES scores different groups with different scores, even when having the same human score.
- Scores over 0.01 suggest unfairness. (Williamson et al. 2012)

4589 annotated essays from 1839 student.

Corpus **Available for** Download:



The DARIUS Corpus was created to develop automated tools for assessing and enhancing German high school students' argumentation skills. It consist of 4.589 essays on "energy" and "automotive" topics from 1.839 students across 33 schools. Students advocated for one of three options on power plants or car engines, produced drafts, received feedback, revised their work, and tackled a transfer task.



Toulmin's

**Pattern** 

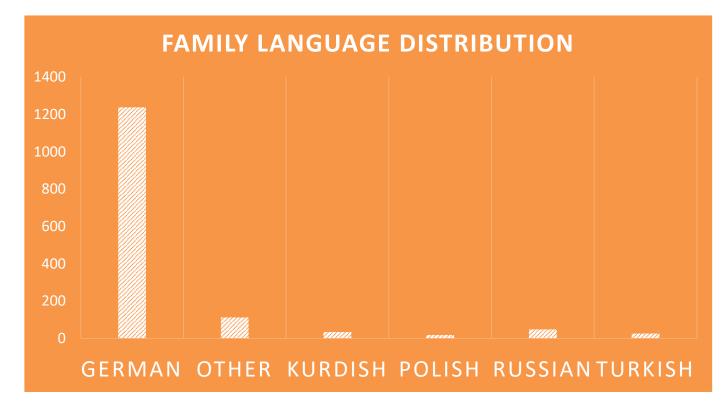
Argumentation

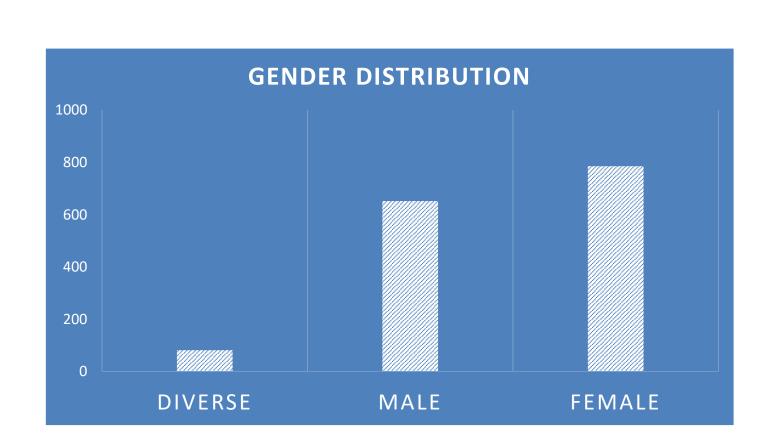
Sequences of

Claim; Data;

argument parts.

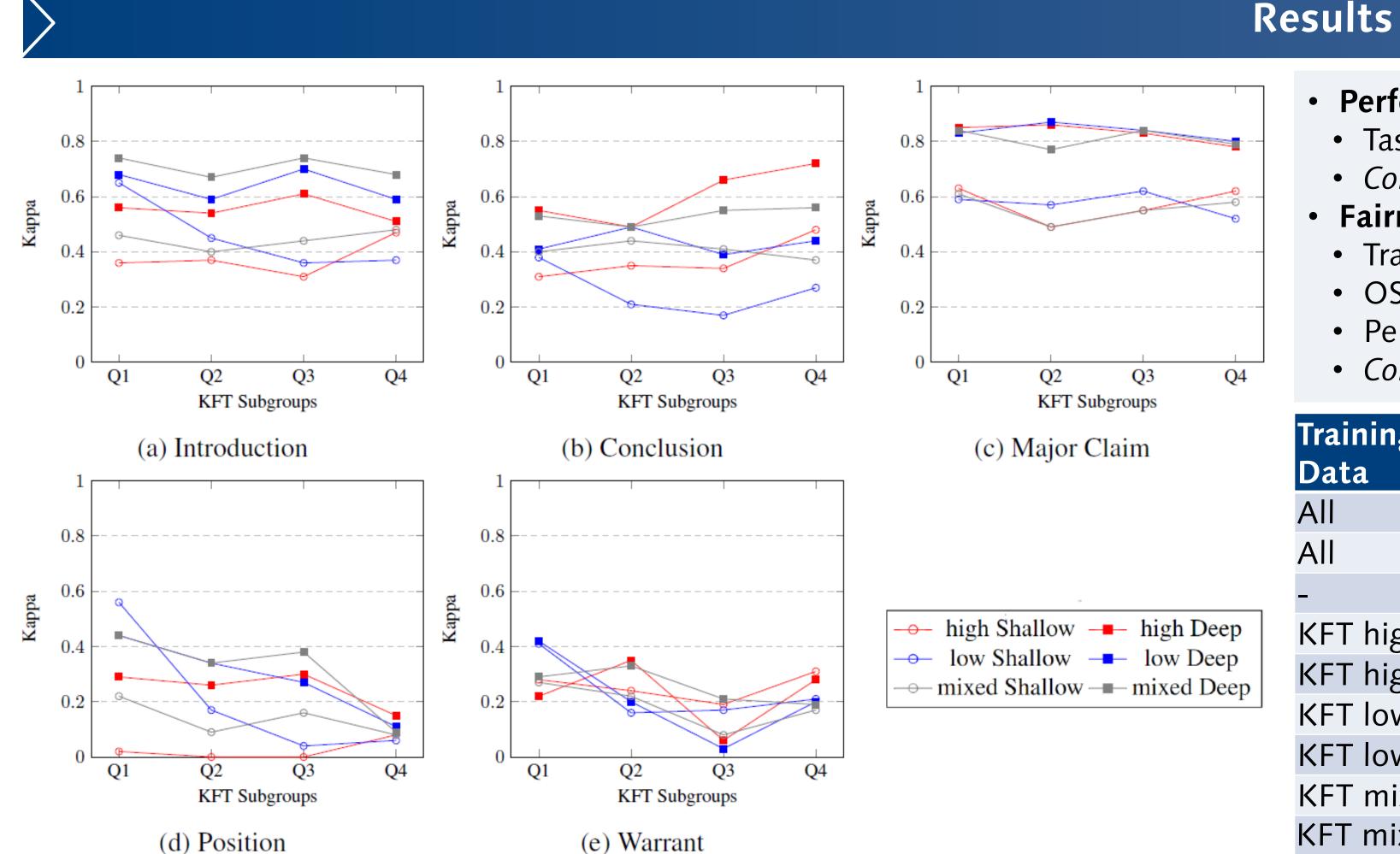
Warrant; Rebuttal.





Fairness and Performance

Evaluated.



#### Performance:

- Task-specific deep learning performs best.
- Considerations: better prompts, other versions etc.
- Fairness:
- Training data from only one student group = lower performance on other groups.
- OSA, OSD, CSD: No values over the threshold of .01.
- Performance and Fairness should both be evaluated.
- Considerations: groups to homogenous? Some groups too small?

Training Data	Model	Intro- duction	Conclusion	Major Claim	Position	Warrant
All	Shallow	0.63	0.55	0.68	0.41	0.43
All	Deep	0.81	0.7	0.88	0.44	0.44
-	LLM	0.60	0.68	0.75	0.32	0
KFT high	Shallow	0.38	0.39	0.57	0.02	0.26
KFT high	Deep	0.56	0.62	0.83	0.29	0.23
KFT low	Shallow	0.47	0.25	0.58	0.37	0.23
KFT low	Deep	0.65	0.44	0.84	0.34	0.2
KFT mixed	Shallow	0.46	0.42	0.56	0.16	0.17
KFT mixed	Deep	0.71	0.54	0.81	0.37	0.25

<sup>1</sup>Leibniz Institute for Science and Mathematics Education at the University of Kiel; <sup>2</sup>CATALPA, FernUniversität in Hagen, Germany; <sup>3</sup>Hildesheim University, Germany





