Deutsche Telekom Stiftung

Finanziert von der Europäischen Union
NextGenerationEU

Leibniz Association

IPN
Leibniz Institute for Science and Mathematics Education

LLM-Feedback zum argumentativen Schreiben

Workshop Große Sprachmodelle in der Bildung

lernen:digital
Kompetenzzentrum
Sprachen/Gesellschaft/Wirtschaft

Ausgangssituation

- Sprache und Schrift als Grundlage für Lernen und gesellschaftliche Teilhabe (Bewerten, Erklären, Reflektieren, Argumentieren; z.B. Freedman et al., 2016; Graham et al., 2020)
- Relevanz qualitativer Übungsmöglichkeiten:
  individuelle Unterstützung durch formative Beurteilung und Feedback (z.B., Graham, 2023)
- Herausfordernd für Lehrkräfte, (regelmäßig) diese zu leisten, insbesondere bei komplexen Leistungen (Zhu et al., 2020)
- Lösungsidee: (Generative) KI als potenzielle Unterstützung der Lehrkräfte (UNESCO, 2023)

Beispiel Aufgabe:

Die Treibhausgasemissionen des Energiesektors sollen drastisch reduziert werden, um die Klimaneutralität zu erreichen. Die Politikerinnen und Politiker eines Bezirks in Norddeutschland fordern daher, den Energiebedarf des Bezirks vollständig durch erneuerbare Energien abzudecken.

Konkret wird beraten, ob hierzu der Bau

- eines Windenergieparks,
- eines Solarparks oder
- einer Wasserkraftanlage

als Energiewandler gefördert werden soll.

Diskutieren Sie die genannten Energiewandler vor dem Hintergrund des Klimawandels auf der Basis von mindestens drei Kriterien aus den Materialien auf der nächsten Seite. Stärken sie Ihre Argumente, indem Sie die Relevanz der Kriterien für ihre Behauptungen erklären.

Beispieltext

Meiner Meinung nach sollte man als Enerdiewandler die Wasserkraftanlage nehmen.

Man sollte dort zwar beachten das es teurer ist als die anderen beiden und auch Hör- und Infraschall hat. Es aber viel Wirkungsvoller ist als der Windpark und der Solarpark. Ebenfalls hat es auch drei mal so viel vom benötigten Jahresertrag den es mindestens haben sollte. Zudem hält es länger wenn nicht sogar doppelt so lange wie die anderen beiden Energiewandler, was auf dauer sinnvoller ist.

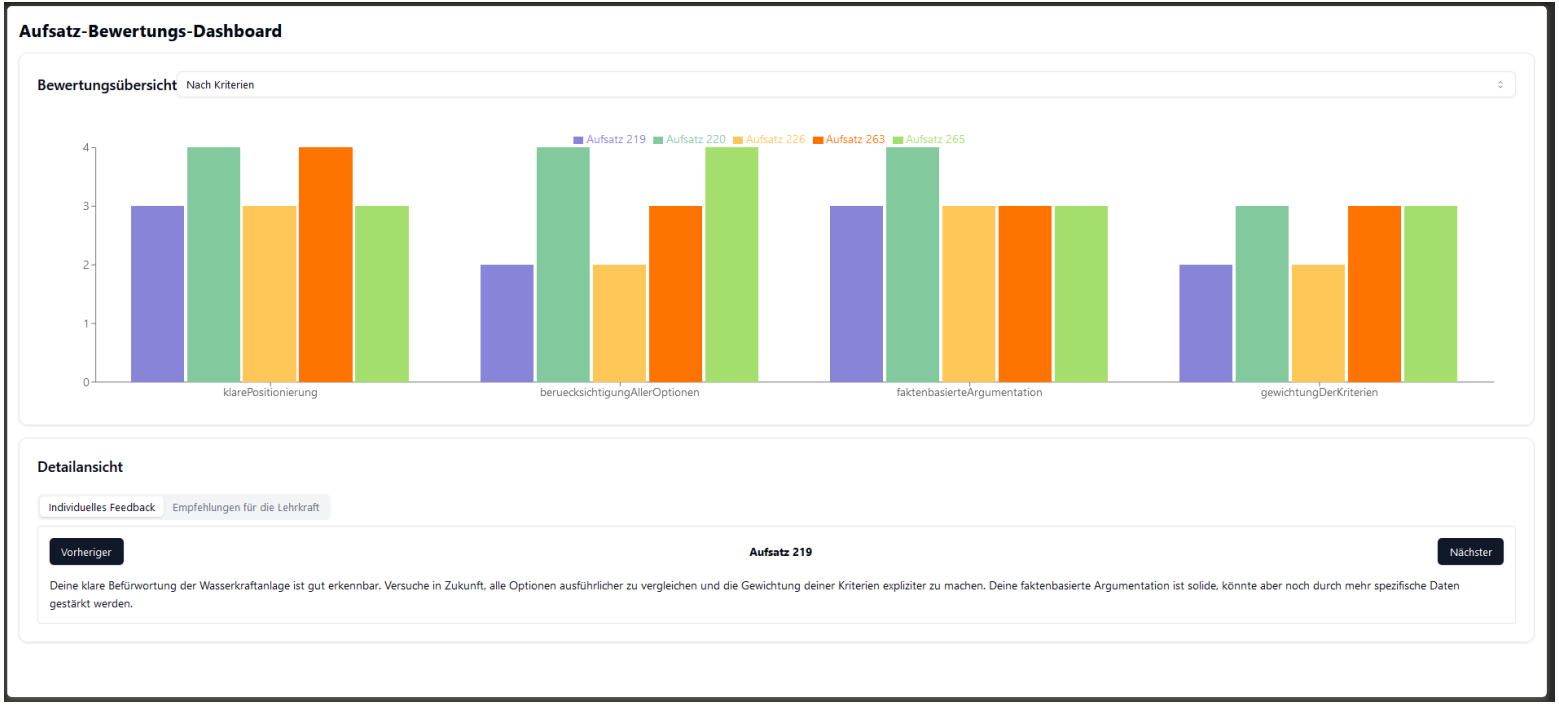
Schreibaufgabe Bewertungskriterien

Hinweise für eine gute Argumentation

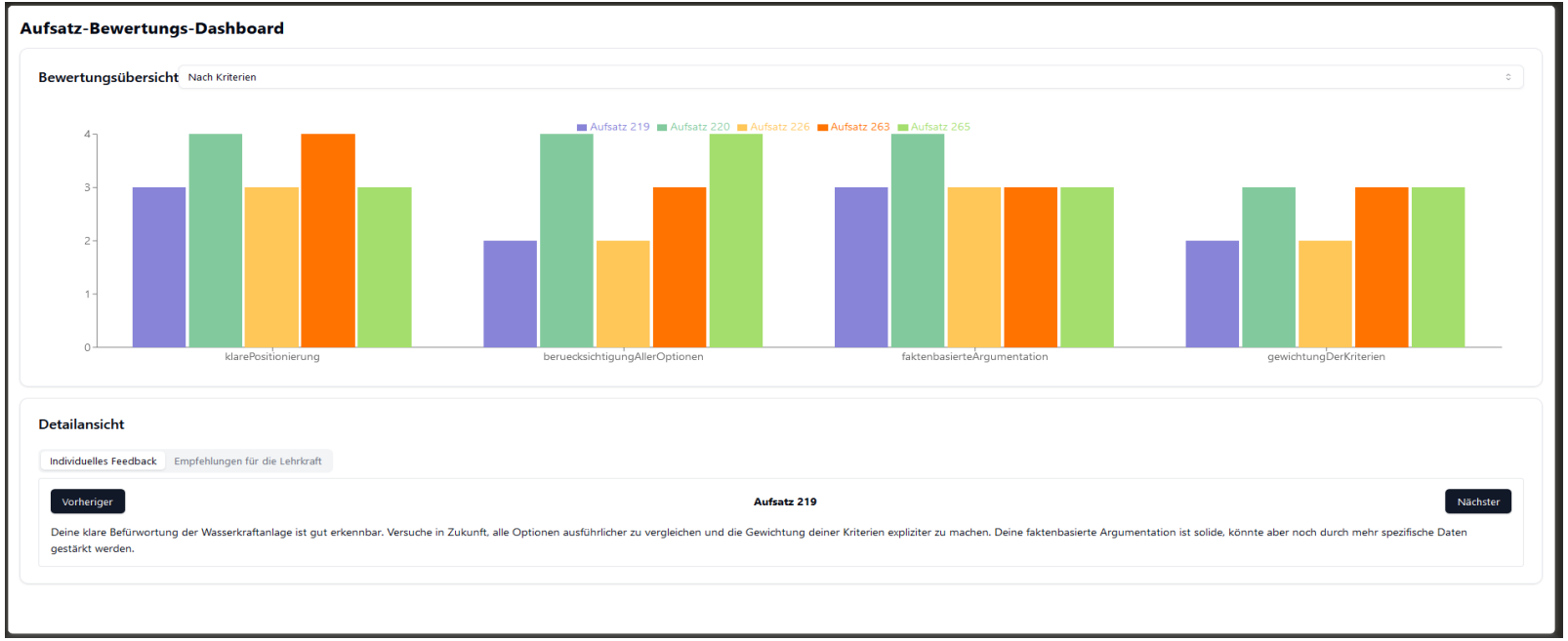
Hinweise:

- eindeutig einen Standpunkt für einen der Energieträger beziehen,
- für jedes von Ihnen genannten Kriterium alle drei Energieträger berücksichtigen,
- die Bedeutung der Fakten durch Begründungen stützen und
- die Bedeutsamkeit der Kriterien für Ihre Entscheidung miteinander vergleichen.

Lehrkräfte-Dashboard



Wie genau ist die automatisierte Beurteilung?



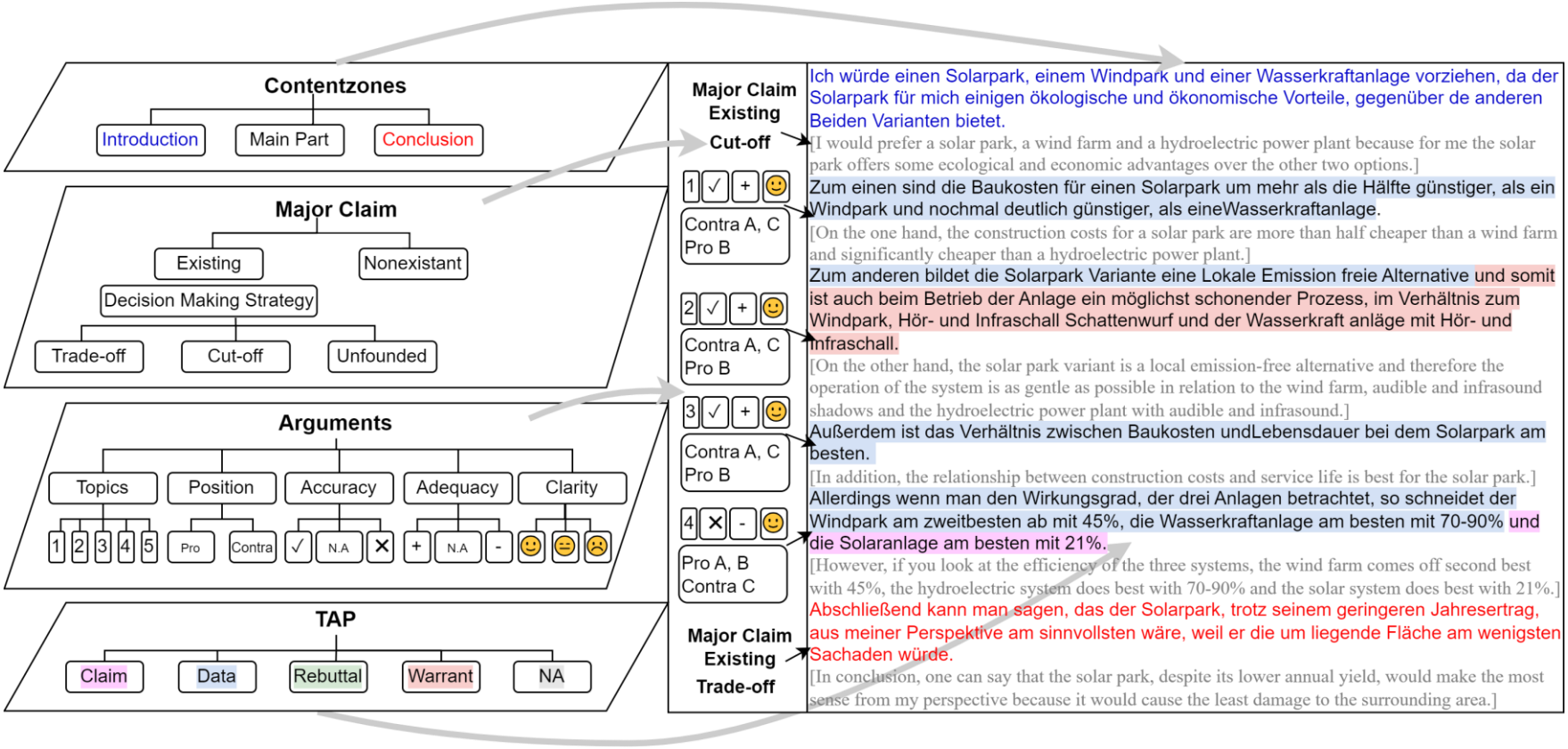
Übersicht Datensatz

- 1839 Teilnehmende aus 114 Klassen von 33 Schulen
  - Alter in Jahren: M = 16.36, SD = 1.47, Bereich = [14, 21], 17.5% fehlend
  - Gender: 42.7% weiblich, 35.5% männlich, 4.4% divers, 17.5% fehlend
  - HISECD: M = 5.13, SD = 1.25, Mdn = 6, Bereich = [0, 6], 23.6% fehlend
- 5225 Texte

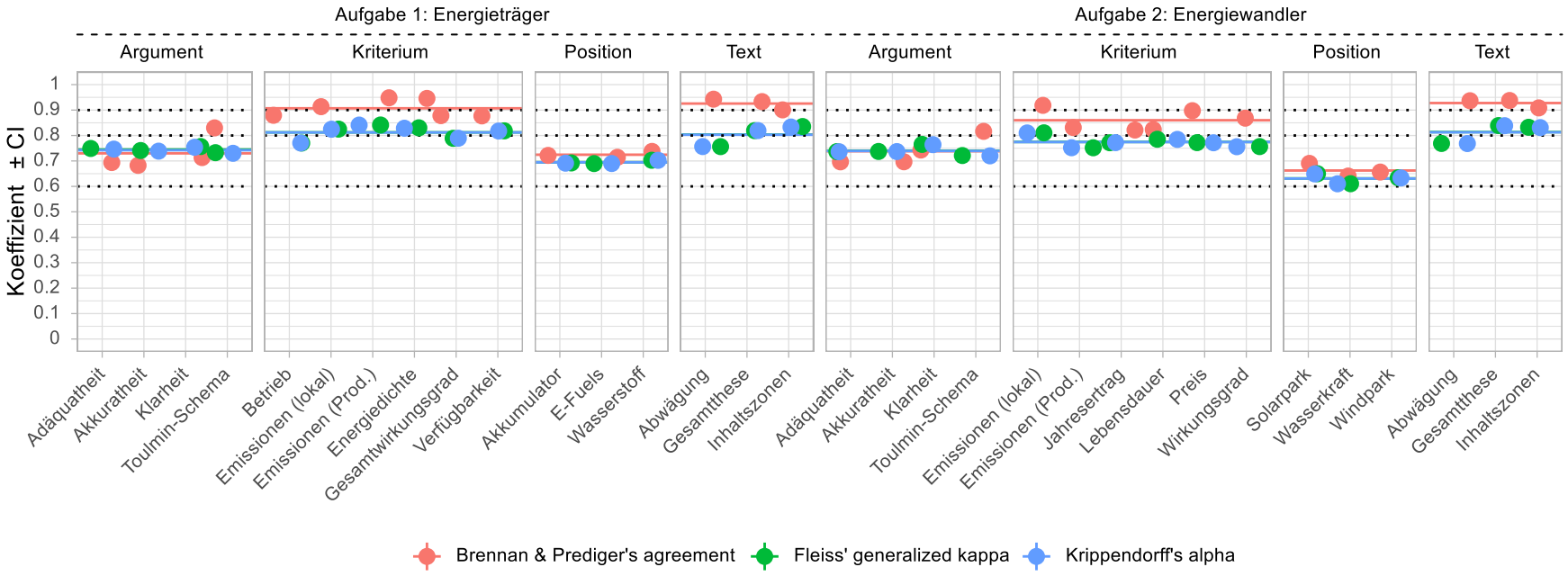
Schultyp	9. Jgst.	10. Jgst.	11. Jgst.	12. Jgst.	13. Jgst.
Gymnasium	369	304	275	259	27
Gemeinschafts-schule	62	89	114	45	34
Berufliche Schulen			121	100	37

Annotation der Texte

- Annotation auf Ebene von Sätzen bzw. Teilsätzen durch 10 Rater
- Jede Person wurde über 4 Wochen trainiert
- Dann über ein Jahr jede zwei Wochen problematische Texte besprochen
- Jedes der über 20 Urteile pro 5000 Texte doppelt kodiert
- Jede Doppelkodierung von dritter Person bei Uneinigkeit kuratiert


Contentzones
Introduction
Main Part
Conclusion
Major Claim
Existing
Nonexistant
Decision Making Strategy
Trade-off
Cut-off
Unfounded
Arguments
Topics
Position
Accuracy
Adequacy
Clarity
1 2 3 4 5
Pro
Contra
✓
N.A
✕
+
N.A
-
TAP
Claim
Data
Rebuttal
Warrant
NA
Major Claim Existing Cut-off
1 ✓ + 🙂
Contra A, C Pro B
2 ✓ + 🙂
Contra A, C Pro B
3 ✓ + 🙂
Contra A, C Pro B
4 ✕ - 🙂
Pro A, B Contra C
Major Claim Existing Trade-off
Ich würde einen Solarpark, einem Windpark und einer Wasserkraftanlage vorziehen, da der Solarpark für mich einigen ökologische und ökonomische Vorteile, gegenüber de anderen Beiden Varianten bietet.
[I would prefer a solar park, a wind farm and a hydroelectric power plant because for me the solar park offers some ecological and economic advantages over the other two options.]
Zum einen sind die Baukosten für einen Solarpark um mehr als die Hälfte günstiger, als ein Windpark und nochmal deutlich günstiger, als eineWasserkraftanlage.
[On the one hand, the construction costs for a solar park are more than half cheaper than a wind farm and significantly cheaper than a hydroelectric power plant.]
Zum anderen bildet die Solarpark Variante eine Lokale Emission freie Alternative und somit ist auch beim Betrieb der Anlage ein möglichst schonender Prozess, im Verhältnis zum Windpark, Hör- und Infraschall Schattenwurf und der Wasserkraft anläge mit Hör- und Infraschall.
[On the other hand, the solar park variant is a local emission-free alternative and therefore the operation of the system is as gentle as possible in relation to the wind farm, audible and infrasound shadows and the hydroelectric power plant with audible and infrasound.]
Außerdem ist das Verhältnis zwischen Baukosten undLebensdauer bei dem Solarpark am besten.
[In addition, the relationship between construction costs and service life is best for the solar park.]
Allerdings wenn man den Wirkungsgrad, der drei Anlagen betrachtet, so schneidet der Windpark am zweitbesten ab mit 45%, die Wasserkraftanlage am besten mit 70-90% und die Solaranlage am besten mit 21%.
[However, if you look at the efficiency of the three systems, the wind farm comes off second best with 45%, the hydroelectric system does best with 70-90% and the solar system does best with 21%.]
Abschließend kann man sagen, das der Solarpark, trotz seinem geringeren Jahresertrag, aus meiner Perspektive am sinnvollsten wäre, weil er die um liegende Fläche am wenigsten Sachaden würde.
[In conclusion, one can say that the solar park, despite its lower annual yield, would make the most sense from my perspective because it would cause the least damage to the surrounding area.]

Interrater-Reliabilität



Ergebnisse: Genauigkeit der automatisierten Beurteilung

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]

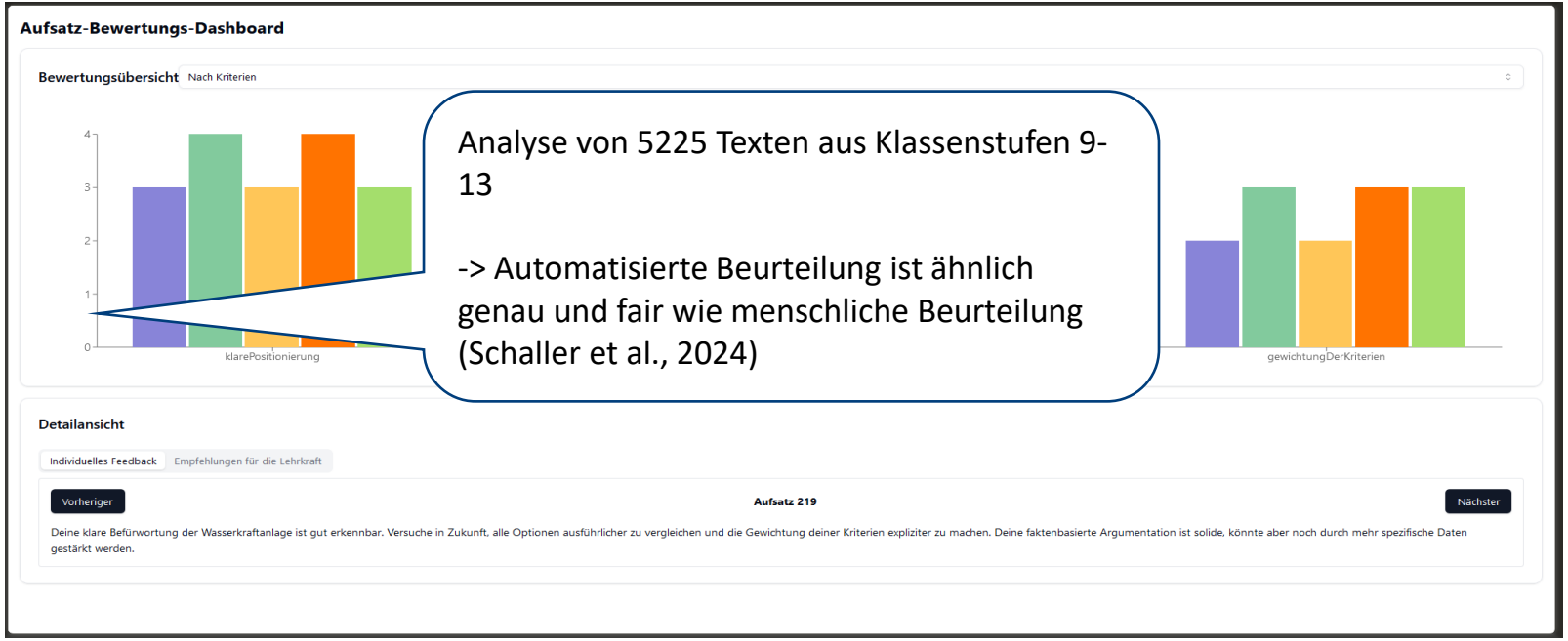
Table 2: Kappa values for the individual classifiers evaluated either on all test essays or on essays from a certain subgroup. We report the minimal and maximal values among the subgroups for each demographic variable.

Ergebnisse

Können Texte genau und fair automatisiert beurteilt werden?

1. Automatisierte Beurteilung ist ähnlich genau und fair wie die Beurteilung des trainierten Beurteilungsteams, auch von ChatGPT
2. Genauigkeit unterscheidet sich wenig zwischen den Gruppen der Schülerinnen und Schüler, wenn diese Gruppe in den Trainingsdaten enthalten ist

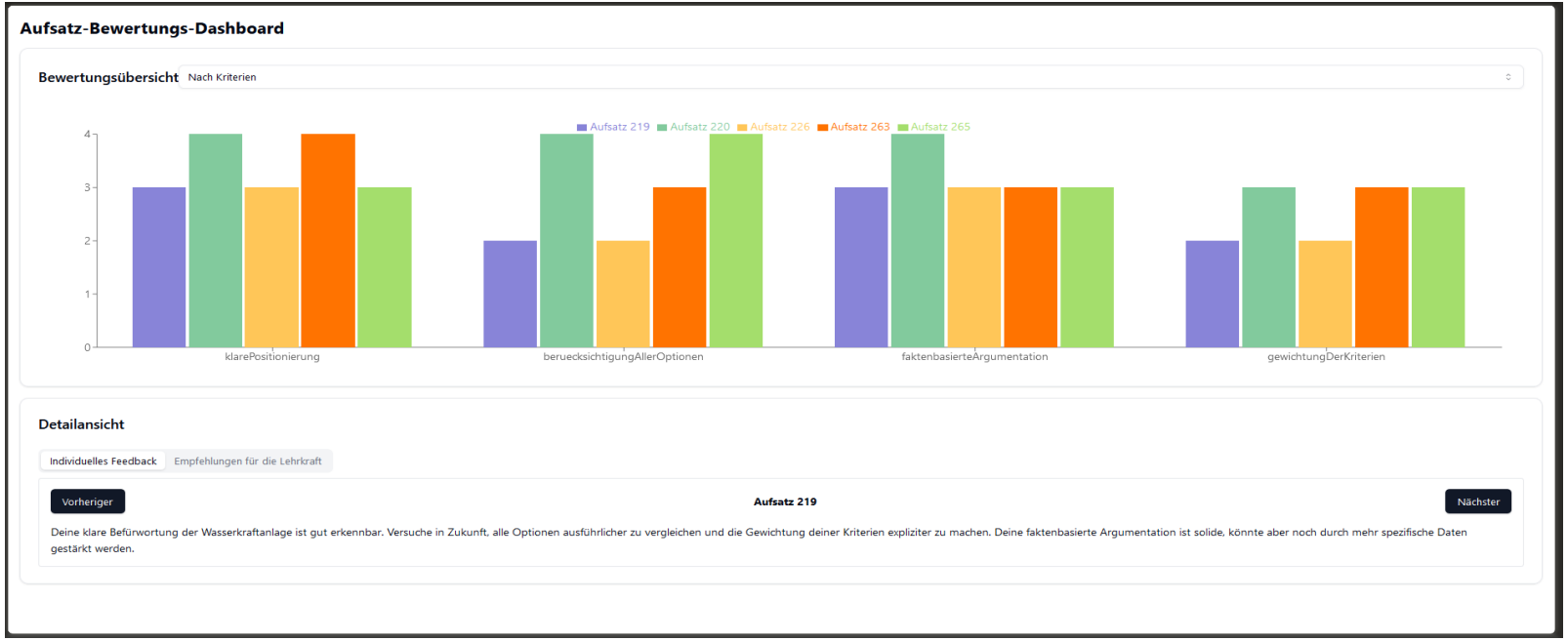
Wie genau ist die automatisierte Beurteilung?



Analyse von 5225 Texten aus Klassenstufen 9-13

-> Automatisierte Beurteilung ist ähnlich genau und fair wie menschliche Beurteilung (Schaller et al., 2024)

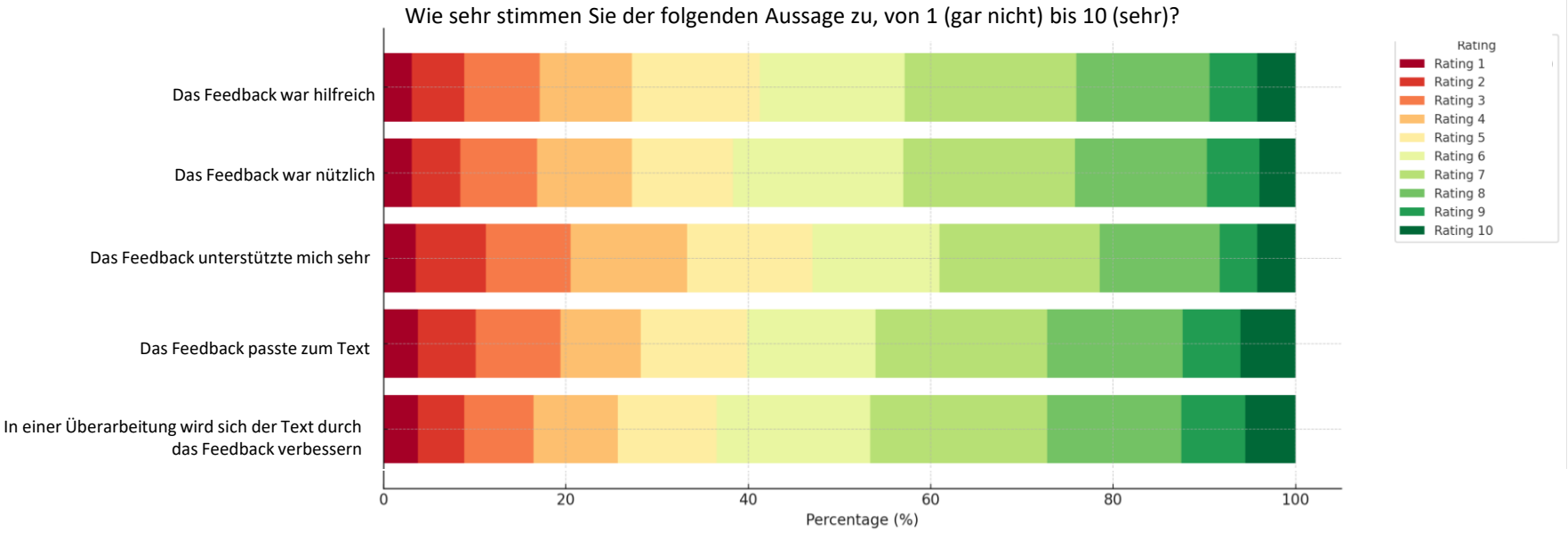
Wie nehmen Lehrkräfte KI-Feedback war?



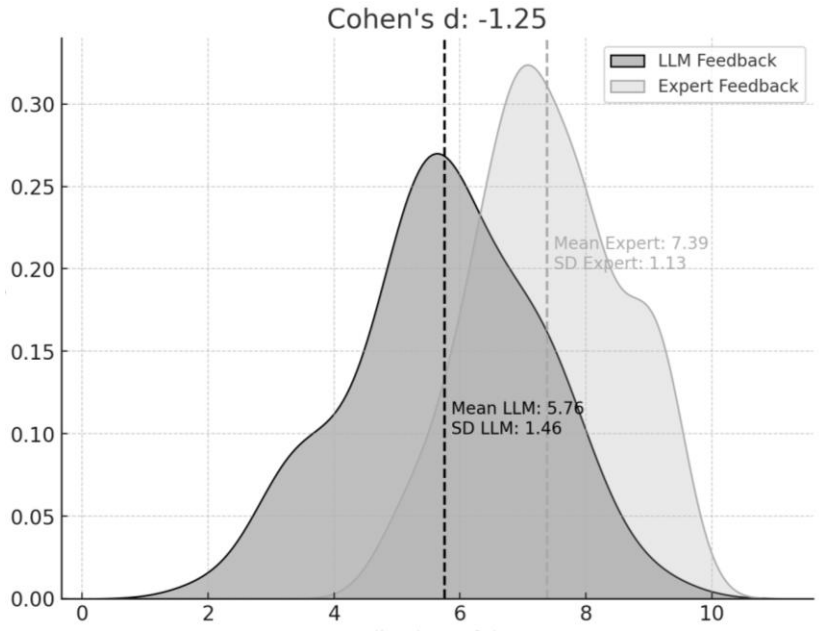
Stichprobe

- Stichprobe: 89 Lehramtsstudierende aller Fachrichtungen, 59 identifizierten sich als weiblich
  - Durchschnittsalter 22 Jahre (SD = 2,84)
  - Durchschnittliche akademische Semester lag bei 4,32 (SD = 3,01).
- Probanden beurteilten je zwei Feedbacks (eins generiert von Experten, eins vom LLM) zu acht Texten
- Nach Beurteilung aller Texte sollten die Probanden die Quelle des Feedbacks zuordnen

Ergebnisse – Feedbackqualität des LLM Feedback



Ergebnisse - Feedbackqualität



Einstellungen von 401 Lehrkräften zu KI-Feedback

Wie wird KI-Feedback im Vergleich mit Lehrkraft-Feedback wahrgenommen, bzw. wie lassen sich die Unterschiede begründen?

The following feedback was created by an **experienced teacher.**

The teacher was instructed to provide structured feedback on 1. content, 2. structure, and 3. style.

Experienced teacher

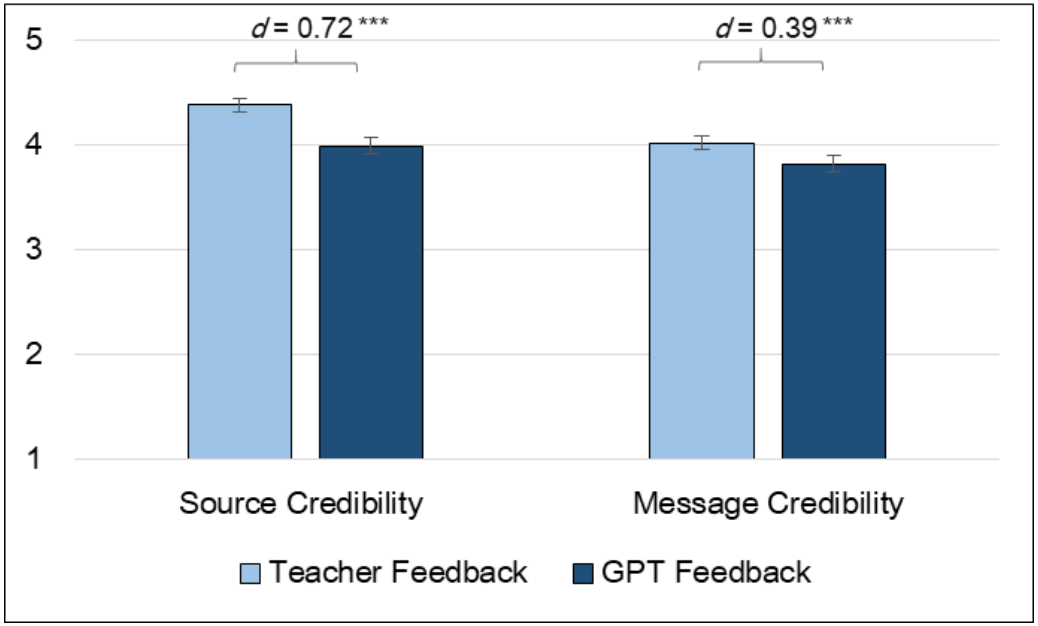
1. Content: In your essay, you address the broad effects of television and social media on concentration, which is a valid and interesting point. However, the focus of the given task is more specific: it's about whether television advertising should be allowed for very young children. To enhance the content of your essay, aim to include arguments related to the marketing tactics used in television ads and how they can affect children aged two to five. Discussing the ethical implications and potential regulations could also strengthen your position.
2. Structure: Your essay format starts with raising the issue, shares personal experiences, and concludes with an opinion, which is good. Nevertheless, for a more effective structure tailored to the task, you might want to consider starting with an introduction that specifically states whether you think advertising to young children should be allowed or not. Following that, present your reasons for and against this, before concluding with your final view. This structure will help you directly address the task and enhance the coherence of your argument.
3. Style: You've written your essay using a personal tone, which makes your writing relatable. For academic assignments, it's beneficial to adopt a more formal tone. Work on using complete sentences and more precise language. Be direct in your statements—for example, instead of saying "its not good for them" you might say "it can negatively impact their development." Don't forget to check your spelling ("parentes" should be "parents", "witch" should be "which", "yout" should be "your") and punctuation, including capitalization ("I cant concentrate" should be "I can't concentrate"). This will greatly improve the clarity and professionalism of your writing.

The following feedback was created by **ChatGPT**:

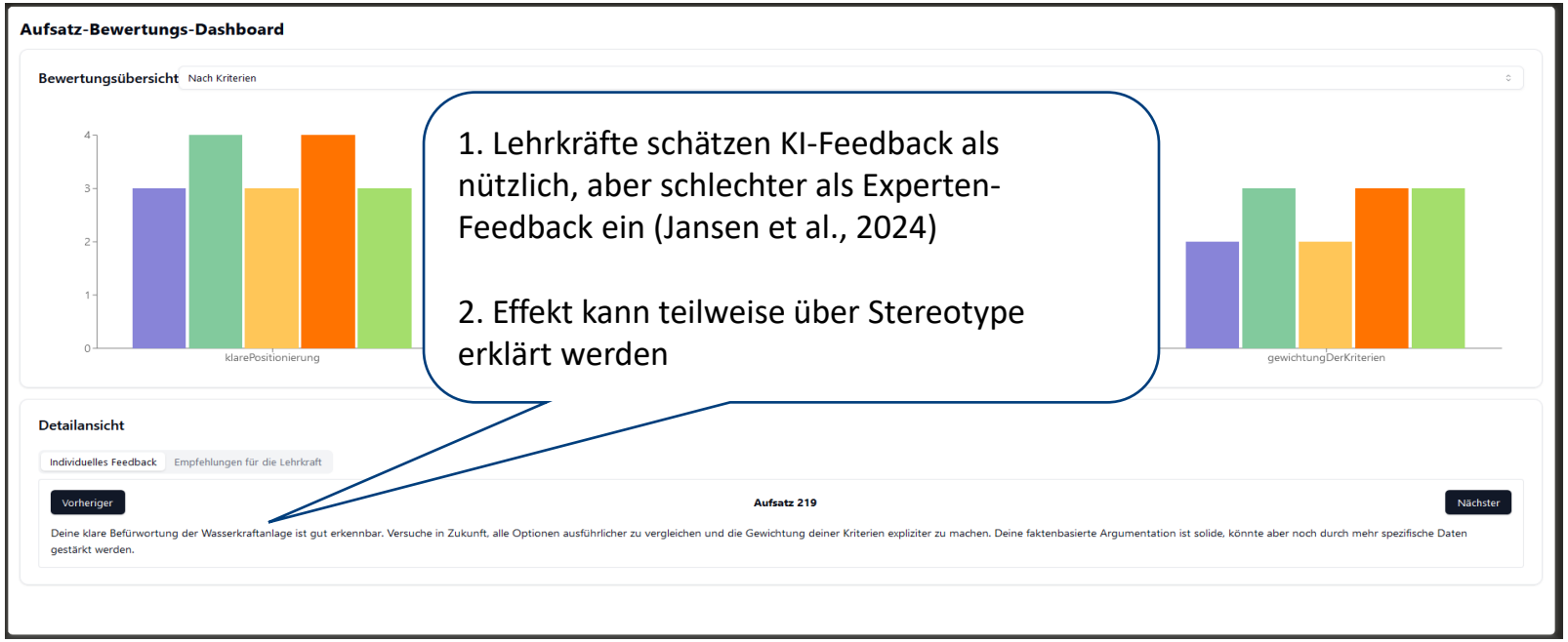
ChatGPT

1. Content: Your essay raises an important topic about the influence of television and social media on young children's ability to concentrate. You provide personal observations, which makes your argument relatable. However, the main task is to discuss whether television advertising specifically should be allowed for children aged two to five. Therefore, it would improve your essay if you focus more on the marketing and advertising aspect of TV, and its impact on children in the specified age group, rather than a generalized discussion on the effects of television and social media.
2. Structure: You've structured your essay with an introduction to the pros and cons of the issue and backed it up with your personal experience. You did a good job in breaking down your thoughts into paragraphs. But for the task given, it's important to clearly introduce the topic, present arguments for both sides, and then provide your personal standpoint in the conclusion. This will make your essay more focused and structured around the question of allowing television advertising for young children.
3. Style: You use a conversational and personal writing style, which is engaging. However, remember that for academic essays, try to maintain a more formal style, using examples and evidence where possible. Also, pay attention to spelling, grammar, and punctuation. Instead of phrases like "really good question" or "its not good," try to use more formal expressions such as "an important consideration" or "it can be detrimental." Also, watch for repetition; you mention concentration several times, so consider varying your examples or focusing on different aspects of the issue.

Ergebnisse - Glaubwürdigkeit



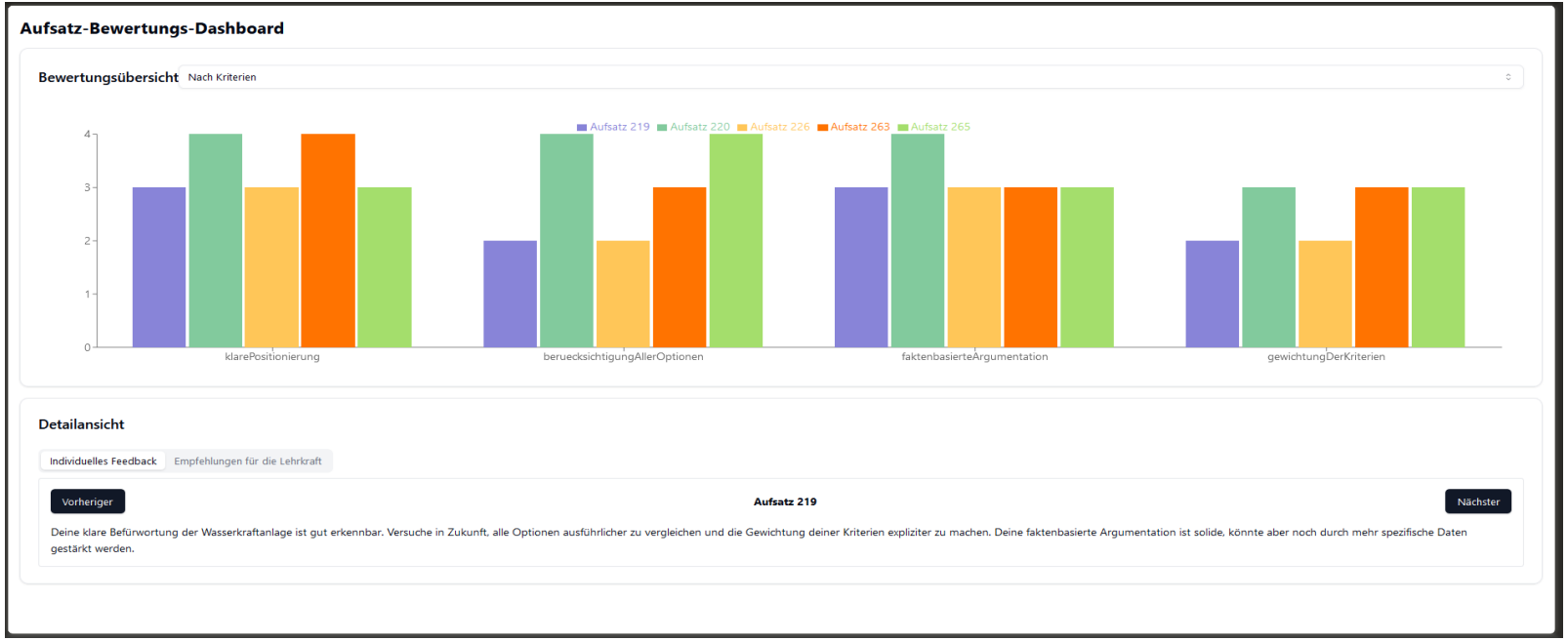
Wie nehmen Lehrkräfte KI-Feedback war?



1. Lehrkräfte schätzen KI-Feedback als nützlich, aber schlechter als Experten-Feedback ein (Jansen et al., 2024)

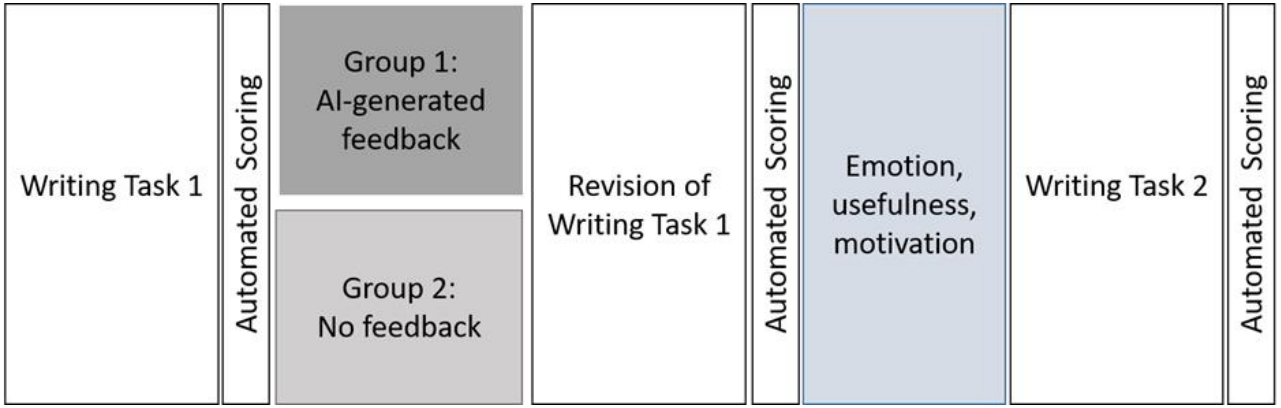
2. Effekt kann teilweise über Stereotype erklärt werden

Profitieren Schüler:innen von KI-Feedback?



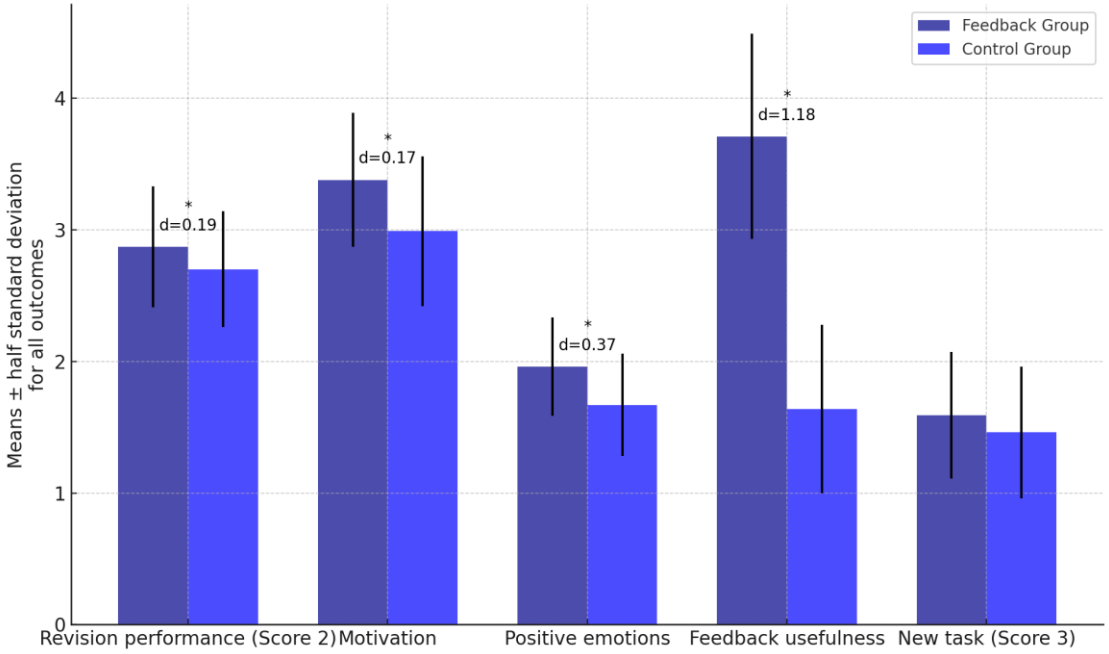
Wie effektiv ist KI-generiertes Feedback für Schüler: innen?

- **Sample**: *N* = 459 (52% weiblich; M_{Age}=16.01; n=203 Feedback, n=256 Kontrollgruppe; 10. Klasse Gymnasium)
- **Design**:



Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

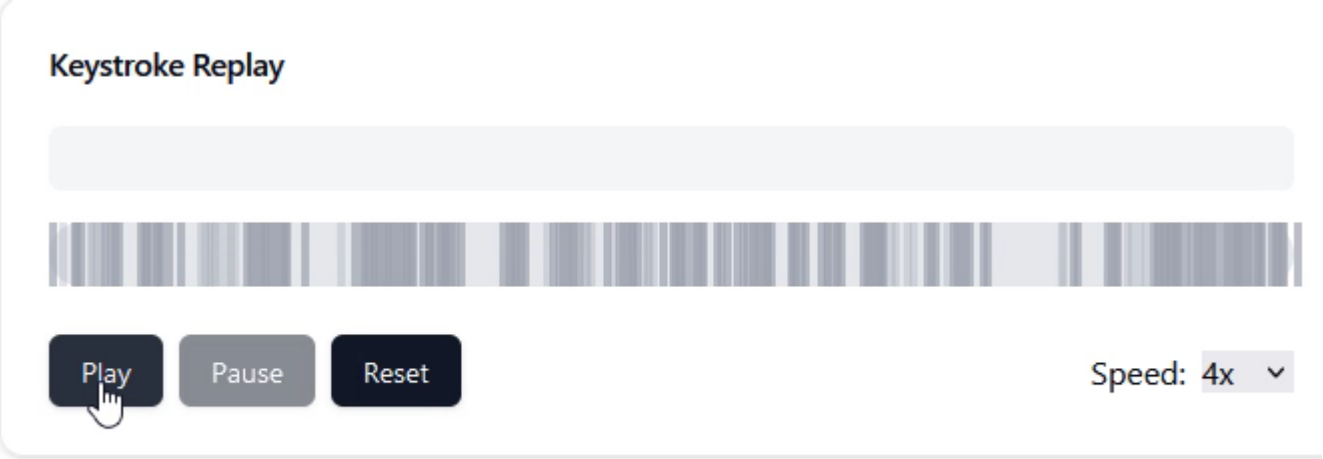
Ergebnisse



Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

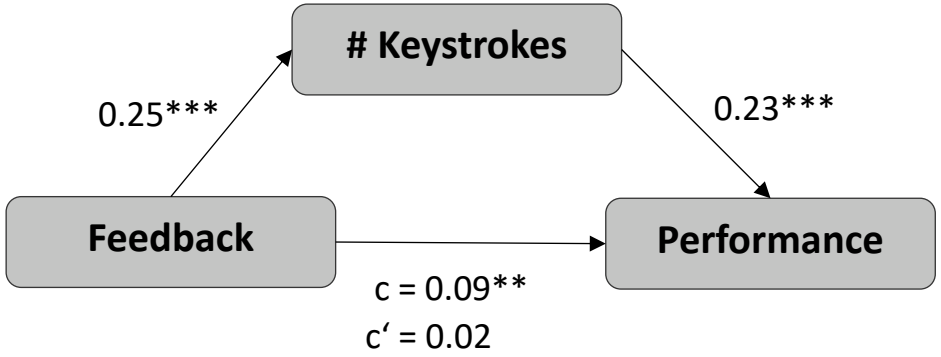
Wie können die positiven Effekte auf die Leistung erklärt werden?

- Idee: Über die Daten beim Schreibprozess
- Time-on-task und Keystrokes



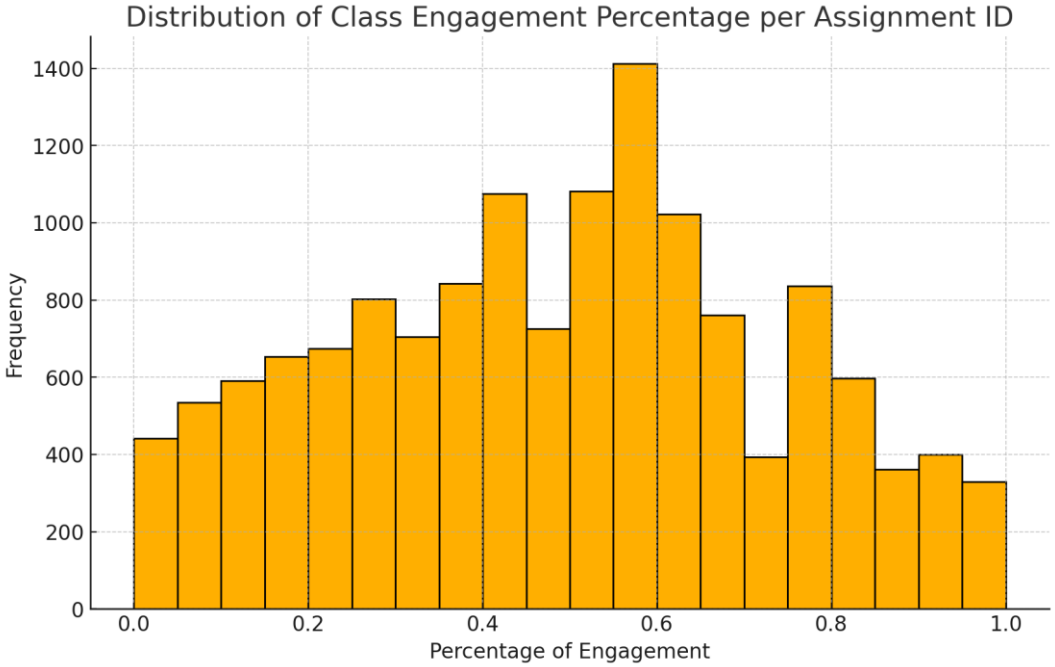
Wie können die positiven Effekte auf die Leistung erklärt werden?

- Idee: Über die Daten beim Schreibprozess
- Time-on-task und Keystrokes (Schiller et al., 2024)

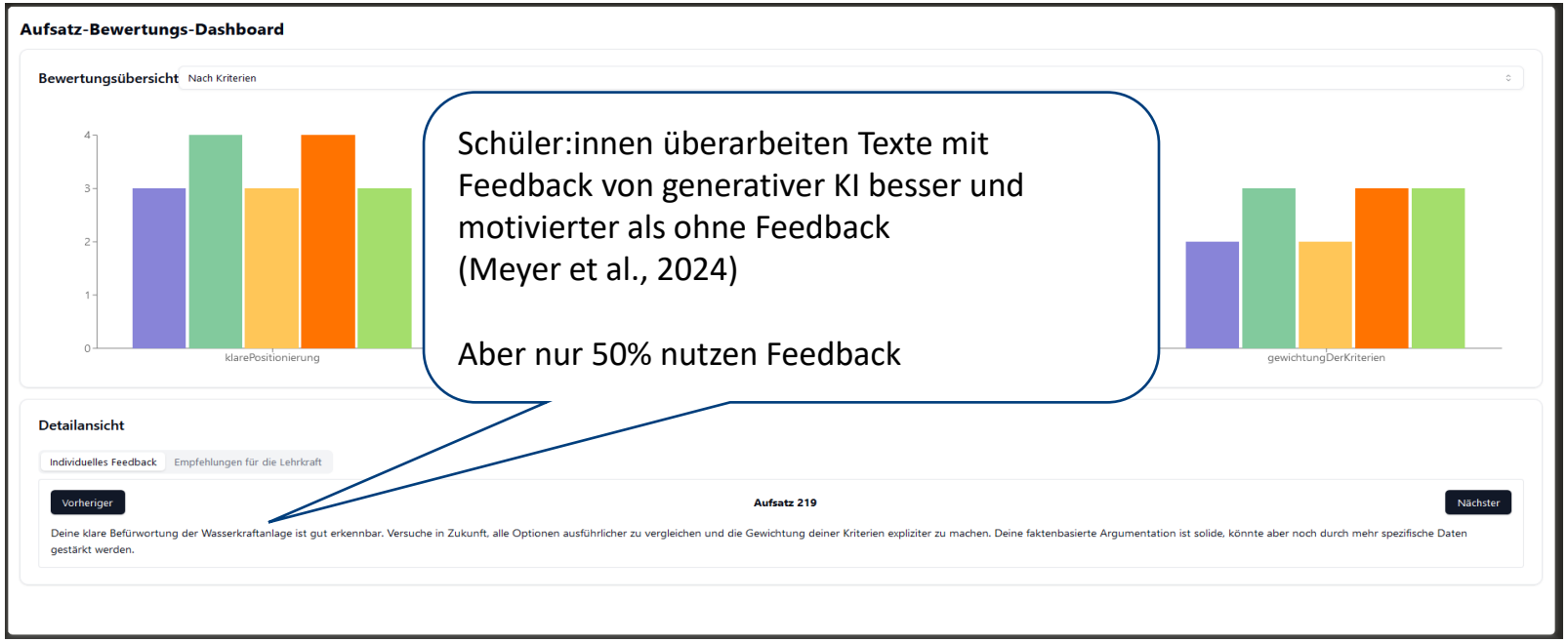


Wie verteilt sich das Engagement in Klassen?

- Häufigkeit Textüberarbeitungen nach generativem KI-Feedback in 655 Schulklassen von der Grundschule bis zur Oberstufe
- Im Mittel ungefähr 50% der 14236 Schüler:innen überarbeiten überhaupt



Profitieren Schüler:innen von KI-Feedback?



Schüler:innen überarbeiten Texte mit Feedback von generativer KI besser und motivierter als ohne Feedback (Meyer et al., 2024)

Aber nur 50% nutzen Feedback

Zusammenfassung

- Automatisierte Beurteilung von Argumentationen, auch von ChatGPT, ist ähnlich genau und fair wie menschliche Beurteilung
- Lehrkräfte beurteilen KI-Feedback schlechter als menschliches Feedback, ein Grund neben Qualitätsunterschieden könnten Stereotype sein
- Feedback generativer KI kann den Schreibprozess unterstützen, da es Engagement der Schreibenden erhöht. Aber nur 50% nutzen es.

Fazit

LLM-Feedback bietet Potenzial Schüler:innen beim argumentativen Schreiben zu unterstützen

Nächstes Ziel ist es, die Bedingungen zu identifizieren, unter denen LLM-Feedback optimal genutzt werden kann:

- Wer profitiert von LLM-Feedback?
- Wie verändert sich der Schreibprozess?
- Wie entwickelt sich die Kompetenz im Vergleich zur Leistung?

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union
NextGenerationEU

Leibniz Association

IPN
Leibniz Institute for Science and Mathematics Education

Vielen Dank!

tjansen@leibniz-ipn.de

lernen:digital
Kompetenzzentrum
Sprachen/Gesellschaft/Wirtschaft

@ipn_Kiel@social.bund.de

@IPN_Kiel

@leibniz_ipn