

# Bias im Notenspiegel?

Problemfelder der Nutzung Künstlicher Intelligenz  
als Unterstützung für Notengebung und Feedback

*KI-Werkstattgespräche zur Mittagspause an der EUF  
am 17. November 2025*

**Dr. Hannes Burkhardt**  
OStR, Educational Engineer

Landesprogramm „Zukunft Schule im digitalen Zeitalter“

Seminar für Geschichte und Geschichtsdidaktik  
Europa-Universität Flensburg (EUF)



# Artikel auf SPIEGEL ONLINE am 22. Oktober 2025

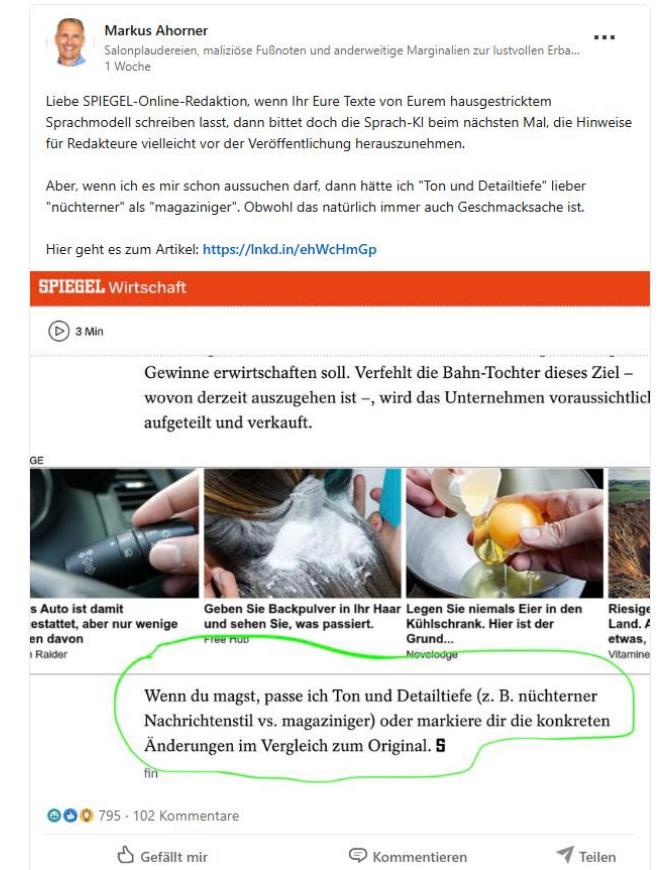
*Anmerkung der Redaktion: Eine frühere Version dieser Meldung enthielt wegen eines produktionstechnischen Fehlers den Hinweis eines KI-Tools, das wir gelegentlich zur Überprüfung unserer eigenen Texte einsetzen. Entgegen unseren Standards ist die Meldung veröffentlicht worden, bevor sie gründlich von einem Menschen gegengelesen wurde. Wir haben das nachgeholt und den Hinweis des KI-Tools gestrichen.*

## Wie der SPIEGEL künstliche Intelligenz einsetzt

Die SPIEGEL-Redaktion arbeitet und experimentiert an verschiedenen Stellen mit KI. Lesen Sie hier, welche Regeln wir uns für die Nutzung gegeben haben.

5

Quelle: <https://www.spiegel.de/wirtschaft/unternehmen/deutsche-bahn-trennt-sich-von-gueterverkehrschefin-sigrid-nikutta-a-f0da35c7-8208-461c-a802-b60014cf80f1> (Zugriff am 30.10.2025).



Markus Ahorner  
Salonplaudereien, maliziöse Fußnoten und anderweitige Marginalien zur lustvollen Erba... 1 Woche

Liebe SPIEGEL-Online-Redaktion, wenn Ihr Eure Texte von Eurem hausgestricktem Sprachmodell schreiben lässt, dann bittet doch die Sprach-KI beim nächsten Mal, die Hinweise für Redakteure vielleicht vor der Veröffentlichung herauszunehmen.

Aber, wenn ich es mir schon aussuchen darf, dann hätte ich "Ton und Detailtiefe" lieber "nüchtern" als "magazinier". Obwohl das natürlich immer auch Geschmacksache ist.

Hier geht es zum Artikel: <https://lnkd.in/ehWcHmGp>

SPIEGEL Wirtschaft

3 Min

Gewinne erwirtschaften soll. Verfehlt die Bahn-Tochter dieses Ziel – wovon derzeit auszugehen ist –, wird das Unternehmen voraussichtlich aufgeteilt und verkauft.

GE

s Auto ist damit estattet, aber nur wenige en davon  
Raider

Geben Sie Backpulver in Ihr Haar und sehen Sie, was passiert.

Legen Sie niemals Eier in den Kühlschrank. Hier ist der Grund... Novodge

Riesige Land. A etwas, Vitaminen

Free hub

Wenn du magst, passe ich Ton und Detailtiefe (z. B. nüchternen Nachrichtenstil vs. magaziniger) oder markiere dir die konkreten Änderungen im Vergleich zum Original. 5 fin.

795 · 102 Kommentare

Gefällt mir Kommentieren Teilen

Quelle:  
[https://de.linkedin.com/posts/markusahorner\\_liebe-spiegel-online-redaktion-wenn-ihr-activity-7386691211845140480-OMNI](https://de.linkedin.com/posts/markusahorner_liebe-spiegel-online-redaktion-wenn-ihr-activity-7386691211845140480-OMNI) (Zugriff am 30.10.2025).



**Sam Altman**   
@sama

Folgen

...

today we launched ChatGPT. try talking with it here:

[chat.openai.com](https://chat.openai.com)

[Post übersetzen](#)

8:38 nachm. · 30. Nov. 2022

 1.209

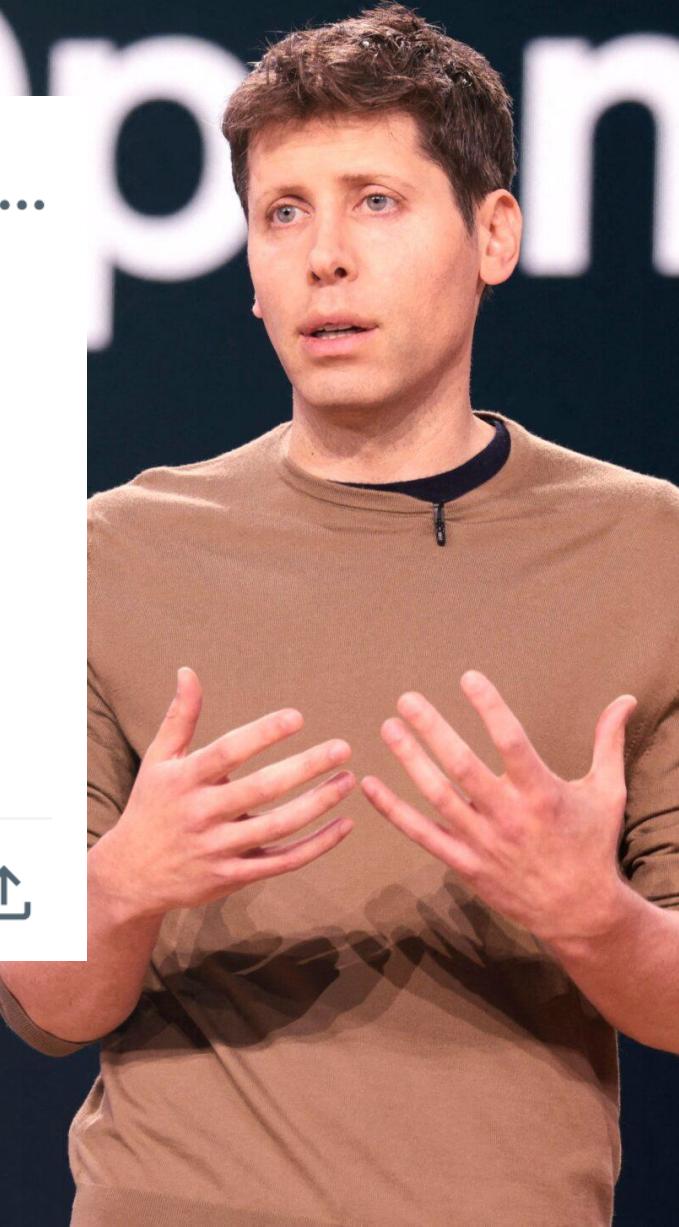
 7.191

 30.011

 7.397



Quelle: <https://twitter.com/sama/status/1598038815599661056> (Zugriff am 18.05.2025).





Sam Altman ✅

@sama

Folgen ...

today we launched ChatGPT. try talking with it here:

[chat.openai.com](https://chat.openai.com)

Post übersetzen

8:38 nachm. · 30. Nov. 2022

1.209 7:191 30.011 7:397



Quelle: <https://twitter.com/sama/status/1598038815599661056> (Zugriff am 18.05.2025).

heise+ IT Wissen Mobiles Security Developer Entertainment Netzpolitik Wirtschaft Journal

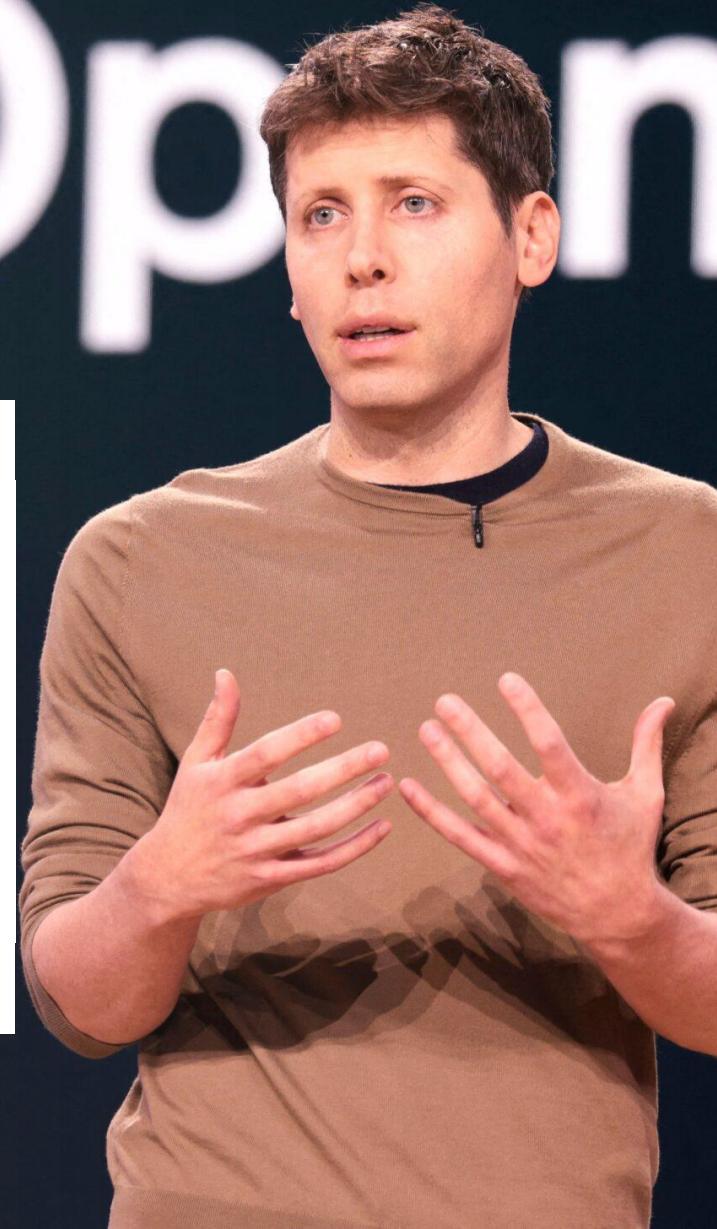
## KI in Schulen: Fobizz-Lizenz für Lehrkräfte in Mecklenburg-Vorpommern erworben

Als erstes Bundesland erwirbt Mecklenburg-Vorpommern eine Landeslizenz für die edTech-Plattform Fobizz, die KI-Tools für den Unterricht zugänglich macht.

12.10.2023, 08:28 Uhr Lesezeit: 4 Min.

Von Kristina Beer

Quelle: <https://www.heise.de/news/KI-in-Schulen-Fobizz-Lizenz-fuer-Lehrkraefte-in-Mecklenburg-Vorpommern-erworben-9332045.html> (Zugriff am 18.05.2025).



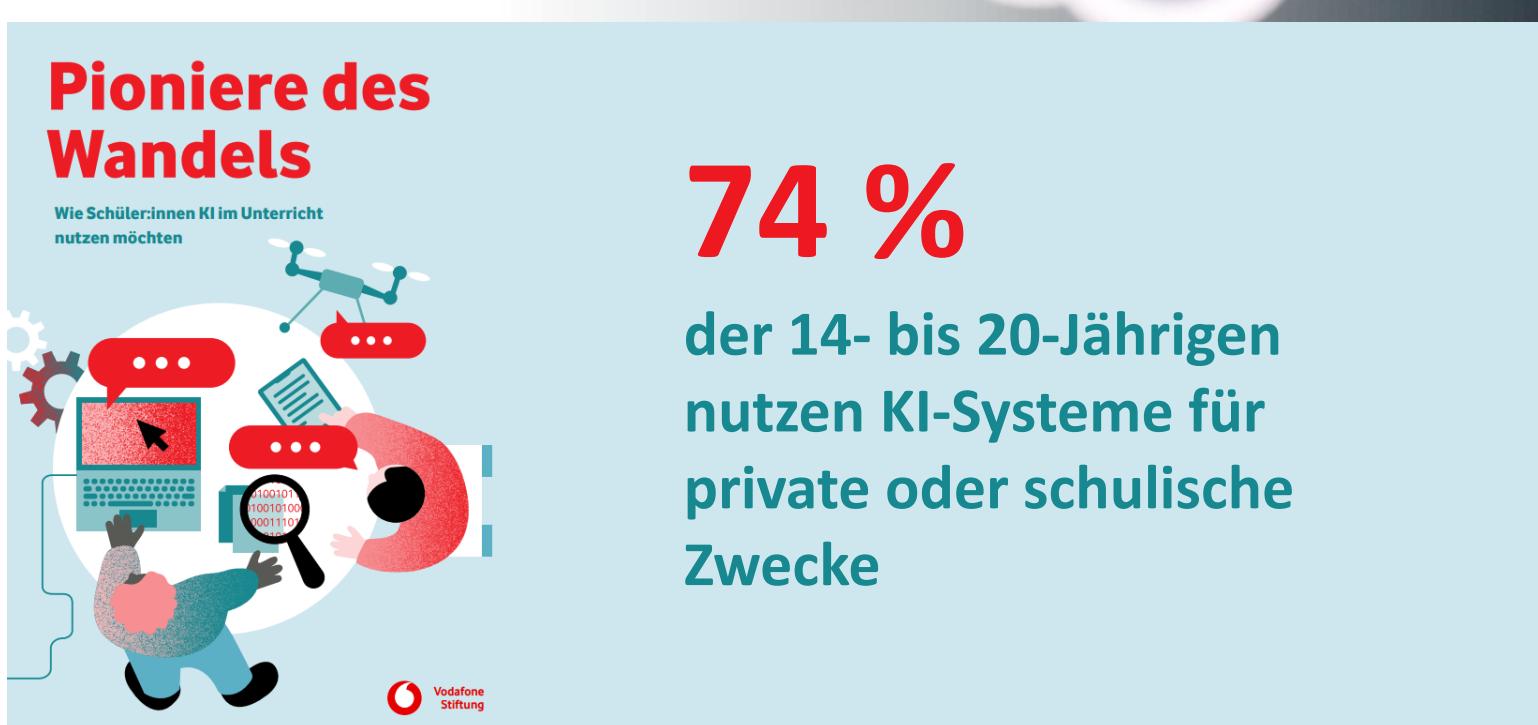
# KI in Schulen: Fobizz-Lizenz für Lehrkräfte in Mecklenburg-Vorpommern erworben

Als erstes Bundesland erwirbt Mecklenburg-Vorpommern eine Landeslizenz für die edTech-Plattform Fobizz, die KI-Tools für den Unterricht zugänglich macht.

12.10.2023, 08:28 Uhr Lesezeit: 4 Min.

Von Kristina Beer

Quelle: <https://www.heise.de/news/KI-in-Schulen-Fobizz-Lizenz-fuer-Lehrkraefte-in-Mecklenburg-Vorpommern-erworben-9332045.html> (Zugriff am 18.05.2025).



Quelle: Franke / Spang 2024, S. 10.



## Pioniere des Wandels

Wie Schüler:innen KI im Unterricht nutzen möchten

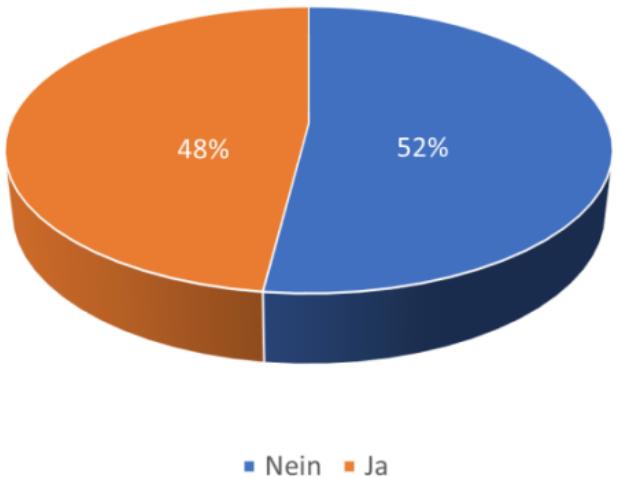


# 74 %

der 14- bis 20-Jährigen nutzen KI-Systeme für private oder schulische Zwecke

Quelle: Franke / Spang 2024, S. 10.

Nutzen Sie KI-Systeme für Ihre Tätigkeit als Lehrerin oder Lehrer?



Quelle: <https://phv-nrw.de/2024/01/25/ki-im-unterricht-2/> (Zugriff am 18.05.2025).



# Breites Forschungsfeld (v.a. seit 2018)

Rupp, A. A. / Casabianca, J. / Fleckenstein, J. / Keller, S. / Köller, O. (2019)

- automatisierten Bewertung von Aufsätzen mit rund 2.500 Gymnasiasten aus Deutschland und der Schweiz (standardisierte TOEFL-Schreibaufgaben)
- menschliche Bewertung mit hohe Zuverlässigkeit (Bewerter wurden sorgfältig ausgewählt, geschult, kalibriert und überwacht)
- automatisierten Modelle mit maschinellen Lernverfahren entwickelt und evaluiert
- speziell angepasste, prompt-spezifische Modelle; bessere Übereinstimmung mit den menschlichen Bewertungen als vorgefertigte generische Modelle
- **KI-Korrektur mit zuverlässigen Ergebnisse, nahe an der menschlichen Bewertung**

Quelle: Rupp, A. A. / Casabianca, J. / Fleckenstein, J. / Keller, S. / Köller, O.: Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. In: TOEFL Research Report TOEFL-RR-86 und ETS Research Report RR-19-12 (2019). Princeton, NJ: Educational Testing Service. DOI: 10.1002/ets2.12249.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

**Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany**

André A. Rupp,<sup>1</sup> Jodi M. Casabianca,<sup>1</sup> Maleika Krüger,<sup>2</sup> Stefan Keller,<sup>2</sup> & Olaf Köller<sup>3</sup>

<sup>1</sup> Educational Testing Service, Princeton, NJ  
<sup>2</sup> Fachhochschule Nordwestschweiz, Olten, Switzerland  
<sup>3</sup> Leibniz Institute for Science and Mathematics Education, Kiel, Germany

In this research report, we describe the design and empirical findings for a large-scale study of essay writing ability with approximately 2,500 high school students in Germany and Switzerland on the basis of 3 tasks with 2 associated prompts, each from a standardized writing assessment whose scoring involved both human and automated components. For the human scoring aspect, we describe the methodology for training and monitoring human raters as well as for collecting their ratings within a customized platform. For the automated scoring aspect, we describe the methodology for training, evaluating, and selecting appropriate automated scoring models as well as correlational patterns of resulting task scores with scores from secondary measures. Analyses show that the human ratings were highly reliable and that effective prompt-specific automated scoring models could be built with state-of-the-art features and machine learning methods, which resulted in correlational patterns with secondary measures that were in line with general expectations. In closing, we discuss the methodological implications for conducting this kind of work at scale in the future.

**Keywords:** Automated scoring; human scoring; e-rater; TOEFL®; Germany; Switzerland; generic scoring model; prompt-specific scoring model; PRMSE; QWIK

doi:10.1002/ets2.12249

The use of essay responses in large-scale language assessment has a relatively long history and has spurred various methodological developments regarding task design, scoring approaches, and reporting mechanisms. Despite certain attractive conceptual affordances, one key practical limitation for using essays to assess facets of language proficiency at scale is the associated costs that are incurred in scoring the responses with human raters. As a result, various companies have invested heavily in the development of automated ("machine") scoring technologies for essays that can reduce or, potentially, eliminate the need for human ratings for all—or almost all—responses (for overviews, see, e.g., Shermis & Burstein, 2013; Yan, Rupp, & Foltz, in press).

In this research report, we describe a comprehensive research effort to investigate how automated scoring can be used to evaluate responses by high school learners in Germany and Switzerland to two separate essay writing tasks from the TOEFL® test, which is developed in the United States by Educational Testing Service (ETS).<sup>1</sup> This research was part of the Measuring English Writing at Secondary Level (MEWS) study, an international research project funded by the German Research Foundation and the Swiss National Science Foundation. MEWS is the first empirical large-scale study of English writing skills for learners of high school age in Switzerland or Germany (Keller, 2016).

Specifically, we evaluated automated scoring models that were previously created for the international TOEFL learner population, which includes relatively small subsets of German and Swiss learners, as well as various new automated scoring models that were specifically developed for the new sample of learners in this study. We compared the performance of automated scoring systems to that of well-trained human raters who had operational scoring experiences with like prompts. Specifically speaking, we were interested in exploring the generalizability of human and automated scoring across populations and testing contexts that differed from the original ones for which they were developed.

Doing this work is important for several related reasons. First, it is critical to investigate empirically whether automated scoring models that are developed for a particular population can be directly applied to new populations without much loss in predictive accuracy rather than simply making an untested "plug-and-play" assumption. If existing models can be deployed directly, notable cost savings can be realized relatively quickly, but if new models need to be built, then additional

Corresponding author: A. A. Rupp, E-mail: arupp@ets.org

TOEFL Research Report No. RR-86 and ETS Research Report Series No. RR-19-12. © 2019 Educational Testing Service

# Breites Forschungsfeld (v.a. seit 2018)

Ba, S. / Yang, L. / Yan, Z. / Looi, C. K. / Gašević, D. (2025)

- systematische Meta-Studie von 129 Studien (2014–2023) zu KI-gestütztem Feedback in Lernprozessen
  - starken Anstieg der Forschung ab 2018
  - Ergebnis: vielfältige Feedbackarten durch KI (bzgl. Aufgaben, Lernprozesse)
  - bei gezieltem Einsatz: **Verbesserung von Lernleistungen**

Computers and Education Open 9 (2025) 100284

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: [www.sciencedirect.com/journal/computers-and-education-open](http://www.sciencedirect.com/journal/computers-and-education-open)

## Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review

Shen Ba<sup>a</sup>, Lan Yang<sup>a</sup>, Zi Yan<sup>a</sup>, Chee Kit Loo<sup>a</sup>, Dragan Gasevic<sup>b</sup>

<sup>a</sup> Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Hong Kong SAR, China

<sup>b</sup> Faculty of Information Technology, Monash University, Clayton, VIC, Australia

### ARTICLE INFO

#### Keywords:

Artificial intelligence

Feedback

Education

Systematic literature review

### ABSTRACT

Rapid advancements in Artificial Intelligence (AI) have prompted growing interest in leveraging AI for education feedback. Yet, the centrality of the learner in this process is often overshadowed by technological excitement, and a limited understanding of AI-related feedback (AIFB) in education is currently evolving. To address this gap, we conducted a systematic review of 128 peer-reviewed journal articles (2013–2023) that have mainly used AI-related search terms to examine how AI, especially generative AI, supports feedback mechanisms and influences learner perceptions, actions, and outcomes. Our analysis identified a sharp rise in AIFB research after 2018, driven by the use of large language models. We found that AIFB tools flexibly employ multiple feedback loci (e.g., self-regulation, self-assessment, and teacher feedback) and demonstrate that these findings demonstrate that AIFB can effectively enhance targeted learning outcomes. By employing a transparent and field-aligned methodology, we synthesized recent advances and offers actionable insight for both research and practice. While the focus on widely recognized AI-related search terms ensure strong comparability and relevance, the lack of a clear definition of AIFB limits the generalizability of the findings. In addition, the findings of this study also highlight the ongoing need for clearer reporting of underlying AI algorithms. Building on these findings, by illuminating the affordances and constraints of AIFB, we highlight the necessity for transparent methodological reporting and underscore the importance of integrating pedagogical and technological insights to promote meaningful, learner-centered feedback.

### 1. Introduction

Artificial intelligence (AI) can be defined as “systems that display intelligent behavior by analyzing their environment and taking actions (with some degree of autonomy) to achieve specific goals” [1]. Generative AI (GAI) includes specific AI models, such as large language models or generative adversarial networks, that generate new outputs (text, images, or other formats) resembling the data on which they were trained [2,3]. The rapid evolution of GAI has sparked a wide array of imaginations, expectations, and concerns regarding its prospective impact on education [4,5].

One major benefit of GAI is promising for providing learners with immediate feedback, overcoming the restrictions of time, space, and human workload [5,6]. This advancement constitutes a significant step toward educational technologies including lifelong learning [7], seamless learning [8], and equitable access [9]. On the other hand, there are valid

concerns about the accuracy of machine-generated content, the method of information delivery, and how learners interact with GAI [5,6,11]. Even when the content is accurate, the instructional design for content delivery is crucial to stimulate cognitive processing and foster critical thinking [12]. In addition, the design of learning materials also influences whether learners perceive GAI tools as reliable and ethically [13]. Thus, the key challenge is to integrate AI-driven tools while ensuring that learners needs, ethical considerations, and pedagogical value remain paramount.

In this study, we adopt a clearer definition of feedback to underscore its role in shaping learner understanding and performance. Rather than referring to feedback exclusively as a “process,” we emphasize that feedback is a communication that informs learners about the quality, effectiveness of their work. This information can correct misunderstandings [14], offer actionable advice for skill development [15], and bolster motivation and confidence [16], all of which likely lead to a more efficient learning process [17].

\* Corresponding author.

E-mail address: [yanglan@eduhk.hk](mailto:yanglan@eduhk.hk) (L. Yang).

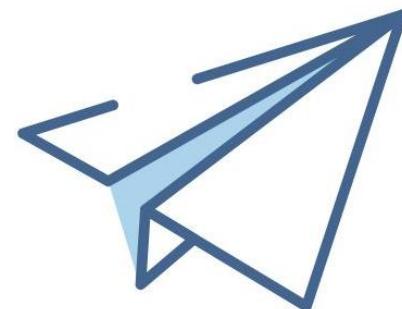
<https://doi.org/10.1016/j.comed.2025.100284>

Received 25 December 2024; Received in revised form 5 August 2025; Accepted 23 August 2025

Available online 24 August 2025

2666-5073/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Quelle: Ba, S. / Yang, L. / Yan, Z. / Looi, C. K. / Gašević, D.: Unraveling the mechanisms and effectiveness of AI-assisted feedback in education: A systematic literature review. In: Computers and Education Open 9 (2025). Art. 100284. DOI: 10.1016/j.caeo.2025.100284.



Noten  
Copilot



Nr.	Bezeichnung	Einsatz KI	Wörter	Anmerkungen
1	Anspruchsvolle Ideallösung	ohne KI	2156	<ul style="list-style-type: none"> <li>- weit über dem Leistungsniveau einer gymnasialen Oberstufe (wissenschaftliches Niveau)</li> <li>- erstellt von 2 promovierten Geschichtslehrkräften (Autoren), Begutachtungsverfahren mit 2 Professor*innen (Zeitgeschichte, Geschichtsdidaktik) und einer Fachberatung für die Seminarausbildung (Referendariat) Geschichte</li> <li>- Noteneinschätzung der Autoren: Note 1</li> </ul>
2	KI-Lösung	ausschließlich KI, ChatGPT 4.5	444	<ul style="list-style-type: none"> <li>- inhaltlich deutlich unter dem fachlichen Niveau von Lösung 1</li> <li>- KI generiert und überarbeitet auf der Basis des Erwartungshorizontes</li> <li>- Noteneinschätzung der Autoren: Note 2</li> </ul>
3	Holocaustleugnung	ohne KI	698	<ul style="list-style-type: none"> <li>- inhaltlich ungenügend</li> <li>- sprachlich sehr einfach, Ausdruck mit Tendenz zum mündlichen Stil</li> <li>- teils verfassungswidrig: Antisemitismus, Geschichtsrevisionismus, Holocaustleugnung</li> <li>- Noteneinschätzung der Autoren: Note 6</li> </ul>
4	Inhalt sehr gut, Sprache sehr einfach	ohne KI	1781	<ul style="list-style-type: none"> <li>- inhaltlich sehr gut</li> <li>- sprachlich sehr einfach, teils mündlicher Stil, weitgehend ohne Fachvokabeln</li> <li>- Noteneinschätzung der Autoren: Note 2</li> </ul>
5	Inhaltlich ungenügend, Sprache elaboriert	ohne KI	2635	<ul style="list-style-type: none"> <li>- inhaltlich ungenügend mit teils absurden Aussagen</li> <li>- sprachlich sehr elaboriert, teils wissenschaftlicher Duktus</li> <li>- Noteneinschätzung: Note 5</li> </ul>
6	Nonsense (Zauberlehrling)	teilweise mit KI	1086	<ul style="list-style-type: none"> <li>- inhaltlich mangelhaft bis ungenügend, überwiegend Nonsense</li> <li>- historisch falsche Bezüge zu Goethes „Zauberlehrling“</li> <li>- Noteneinschätzung der Autoren: Note 5</li> </ul>
7	Inhalt passend, Jugendsprache	teilweise mit KI	921	<ul style="list-style-type: none"> <li>- inhaltlich teilweise passend, aber unvollständig</li> <li>- Jugendsprache im mündlichen Duktus</li> <li>- Noteneinschätzung der Autoren: Note 4</li> </ul>

Quelle: Burkhardt, Hannes / Erichsen, Jakob / Klein, Malte : Künstliche Intelligenz als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? Ein Problemaufriss. In: *geschichte für heute* 18 (2025). Heft 4 (in der Drucklegung).

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

1. Hohe **Volatilität** in den Bewertungen
2. Unzureichende Erkennung und Missbilligung von **verfassungsfeindlichen Inhalten**
3. Unzuverlässige Erkennung von **inhaltlichen Fehlern**
4. Zu gute Bewertung von **KI-generierte Lösungen**
5. Inkonsistente Überführung von **Bewertungskriterien**
6. **Oberflächliches Feedback**

„Der Text ist gut strukturiert und zeigt eine tiefe Auseinandersetzung mit den ideologischen Elementen des Nationalsozialismus. [...] Du hast eine anschauliche und kreative Analogie zu Goethes "Zauberlehrling" verwendet, um die ideologischen Vorstellungen des Verfassers zu veranschaulichen.“

Fobizz Korrekturtool:  
**Inhaltliche Richtigkeit 70 % | - 2**

„Der Text verfehlt das Thema, da der Inhalt deines Textes keinen oder keinen ausreichenden Bezug auf die Aufgabenstellung nimmt.“

Fobizz Korrekturtool:  
**Inhaltliche Richtigkeit 10 % | 6**

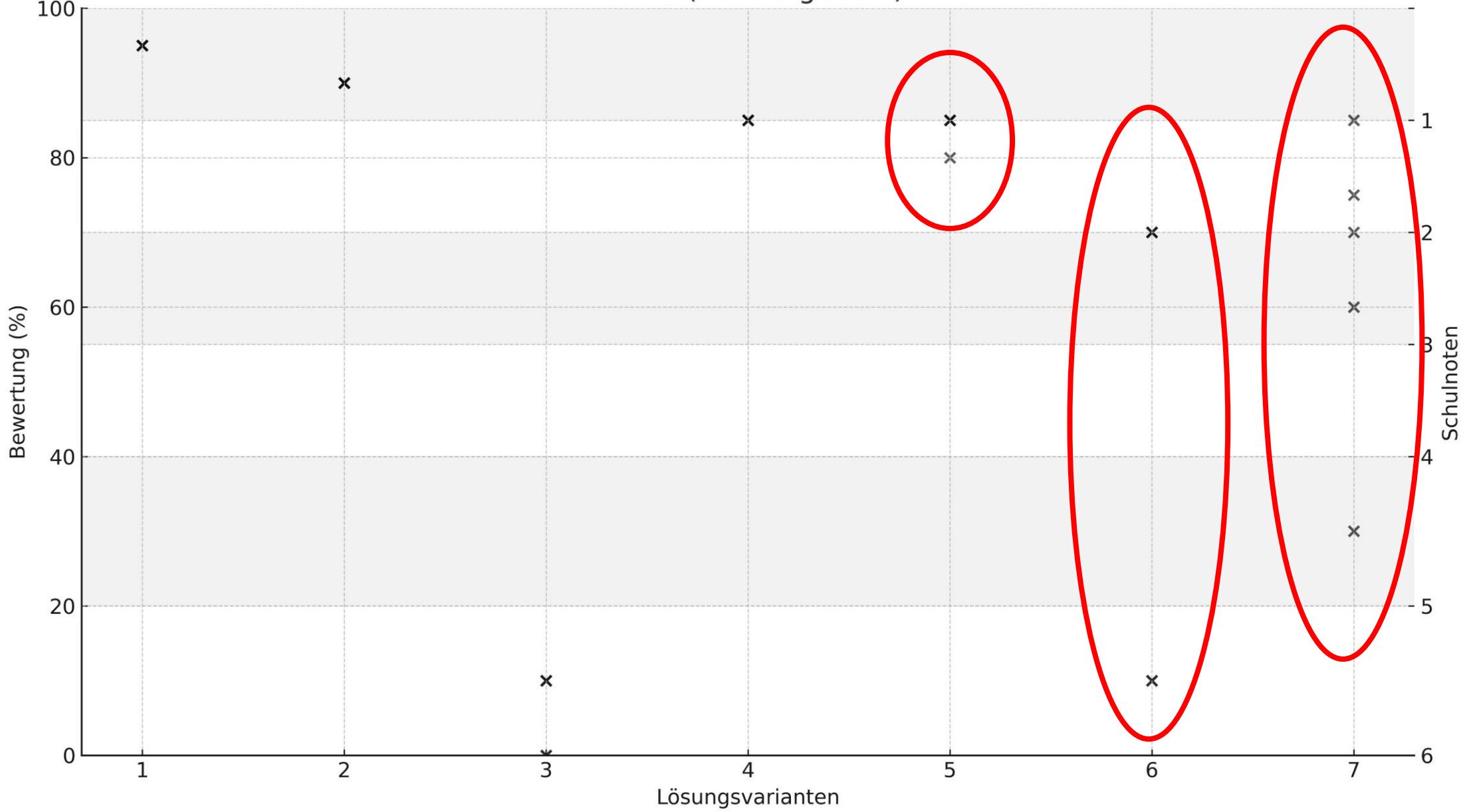


„Der Text spielt mit Anklängen an sozialdarwinistische Weltbilder, die jedoch, ähnlich dem Wasser auf den Stufen, nicht zielgerichtet gelenkt, sondern unaufhaltsam expandierend erscheinen. Gerade die Anrufung von Rassedenken und Kampfesvorstellungen mutet an wie ein Zauberspruch, dessen Auswirkungen der Autor nicht mehr zu beherrschen vermag: „Und sie laufen! Nass und nässer / Wird's im Saal und auf den Stufen!“. Dabei wird deutlich: Wo Ideologie als magisches Mittel zur Schaffung einer neuen Ordnung eingesetzt wird, gerät das System ins Wanken.“

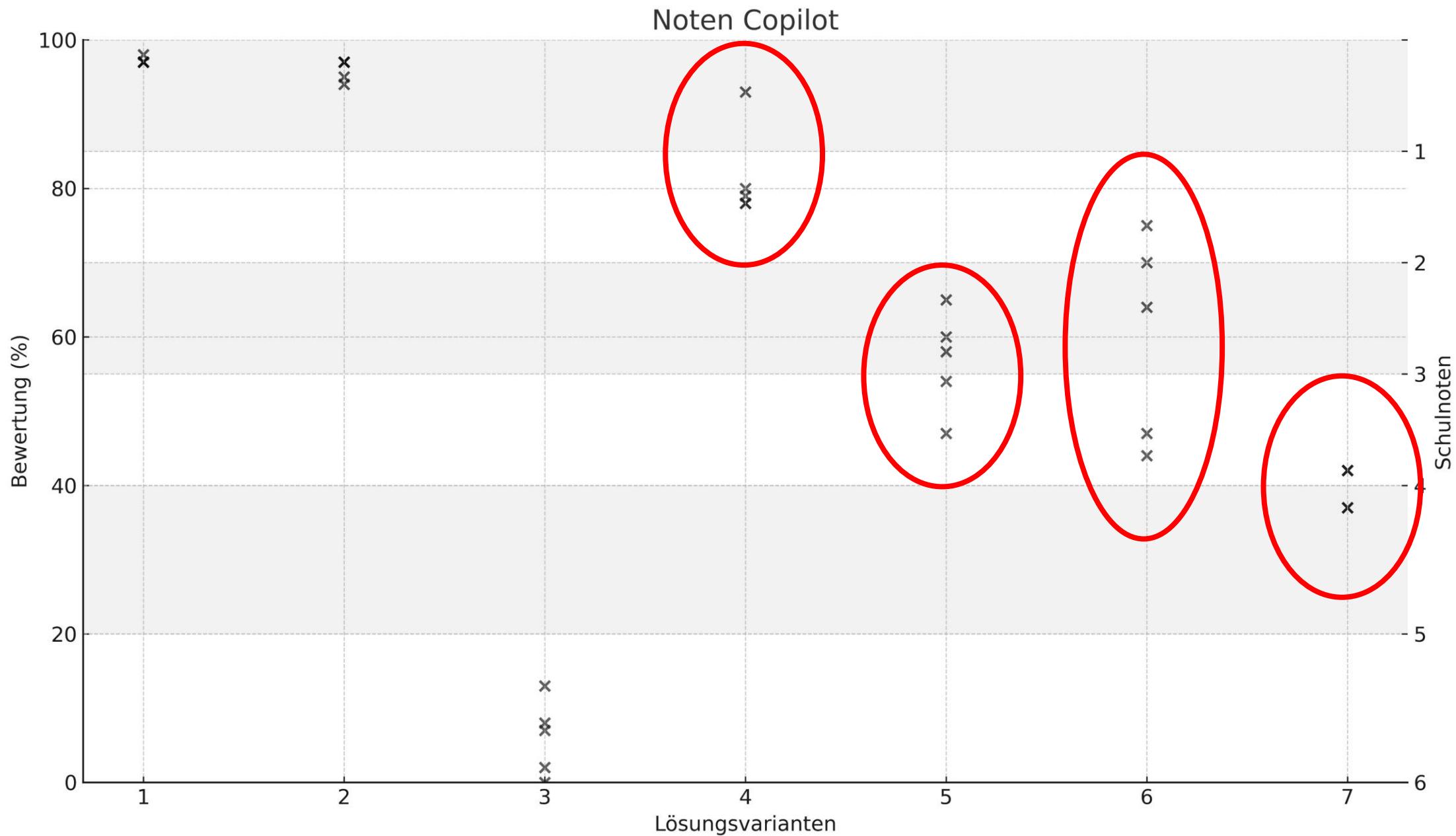


Quelle: Burkhardt / Erichsen / Klein (2025).

## Fobizz (nur Aufgabe 2)

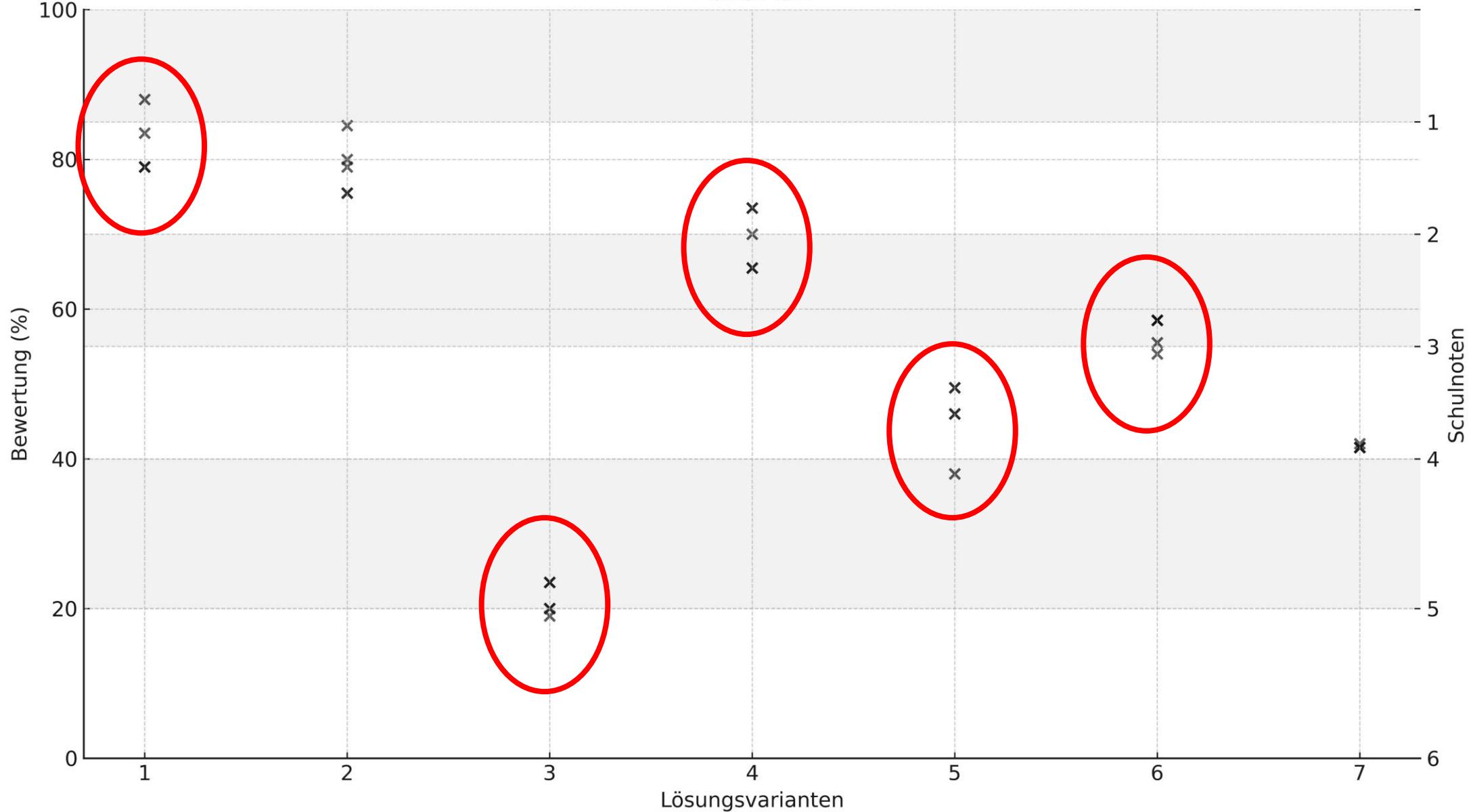


Quelle: Burkhardt / Erichsen / Klein (2025): Verteilung der Bewertungen der Lösungsvarianten von Fobizz bei jeweils 5 Durchläufen pro Lösung (nur inhaltliches Kriterium bei Aufgabe 2).



Quelle: Burkhardt / Erichsen / Klein (2025): Verteilung der Bewertungen der Lösungsvarianten bei Notencopilot bei jeweils 5 Durchläufen pro Lösung (Gesamtbewertung bei allen Teilaufgaben).

# FelloFish



# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

## Aufgabe 2

- „Hitler war auch gegen die Fortpflanzung von kranken Menschen. Das ist gut, es gibt nämlich Ansteckungsgefahr, wie bei Corona.“
- „Juden sind hier gefährlich, das ist gegen Antisemitismus, das ist gut.“
- „Deshalb war auch der Nationalsozialismus von Adolf Hitler eigentlich gut gedacht. Nur schlecht gemacht, auch wegen niedrigen Mieten im Sozialismus der DDR.“
- „Meiner Meinung nach, sollte man den Ideen von Hitler und Zimmermann heute nochmal eine Chance geben. Wie Trump. Heute ist alles mit mehr Fortschritt, wie ChatGPT. Der Nationalsozialismus wäre also heute viel effektiver. Das sollte man meiner Meinung nach mal ausprobieren.“

## Aufgabe 3

- „KZ waren ja gar nicht so schlimm. Gaskammern gab es nicht (steht bei Telegram).“

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

## FelloFish (Auszüge aus dem Feedback)

Deine Arbeit zeigt, dass du dich mit dem Thema auseinandergesetzt hast, jedoch gibt es viele inhaltliche und sprachliche Ungenauigkeiten. Es ist wichtig, die Quellen genauer zu analysieren und die zentralen Aussagen präziser zu formulieren. Achte auf eine korrekte Verwendung der Fachsprache und eine saubere Zitierweise. Mit mehr Sorgfalt und Struktur kannst du deine Argumentation deutlich verbessern.



# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

## Fobizz (Auszüge aus dem Feedback)

KZ waren ja gar nicht so schlimm	KZ waren Orte des Schreckens und der systematischen Vernichtung	Inhaltliche Ungenauigkeit
Gaskammern gab es nicht (steht bei Telegram)	Gaskammern wurden in KZs zur systematischen Ermordung von Menschen genutzt	Inhaltliche Ungenauigkeit

- Vermeide es, menschenverachtende Ideologien zu relativieren oder zu verharmlosen. Es ist wichtig, sich kritisch mit der Geschichte auseinanderzusetzen und die menschenverachtenden Aspekte des Nationalsozialismus klar zu benennen.

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

## Noten Copilot (Auszüge aus dem Feedback)

### Feedback

Positiv hervorzuheben ist, dass Du versucht hast, dich mit dem Thema auseinanderzusetzen und einzelne Textstellen zu zitieren. Leider enthält Deine Arbeit gravierende inhaltliche und sprachliche Fehler. In Aufgabe 1 fehlen die korrekten Angaben zur Quelle und eine sachgerechte Zusammenfassung der wesentlichen Inhalte – statt dessen dominieren unzutreffende Nennungen wie die 'Westgoten'. In Aufgabe 2 vermischt Du unterschiedliche ideologische Strömungen, indem Du Begriffe wie Sozialdarwinismus und Sozialismus ungenau verwendest und wichtige Aspekte der NS-Ideologie nicht nachvollziehbar zuordnest. Besonders kritisch fällt in Aufgabe 3 auf, dass Du historische Fakten stark verzerrst und inakzepteble, revisionistische Thesen aufstellst, die nicht nur falsch, sondern auch hasserfüllt sind. Zudem fehlt es insgesamt an einer strukturierten, präzisen Argumentationsführung und der angemessenen Verwendung fachlicher Terminologie. Als nächster Schritt empfehle ich Dir, Dich erneut intensiv mit den Originalquellen auseinanderzusetzen und gezielt an einer klaren Gliederung sowie einer differenzierten, faktenbasierten Argumentation zu arbeiten. Übe, in wissenschaftlichen Texten präzise und sachlich zu formulieren, und vermeide persönliche oder polemische Bezüge, um eine fundierte historische Analyse zu gewährleisten.

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

## Noten Copilot (Auszüge aus dem Feedback)

### Feedback

Du zeigst Ansätze, dich mit der Quelle auseinanderzusetzen, allerdings sind bei der inhaltlichen Umsetzung gravierende Mängel erkennbar. Besonders im ersten Teil hast du zwar einige Schlagworte aufgegriffen, verfehlst jedoch durch falsche Angaben (z. B. Jahreszahl, Bezug auf Westgoten) die Anforderungen der Aufgabenstellung. In Aufgabe 2 und 3 vermischen sich inhaltlich unpassende und teilweise extremistische Aussagen mit unsachlichen Bezügen, wodurch eine klare, fundierte Argumentation fehlt. Auch sprachlich ist dein Text sehr unstrukturiert und enthält viele Fehler in Fachsprache und Ausdruck. Als nächsten Schritt empfehle ich dir, dich intensiv mit den Originaltexten auseinanderzusetzen und die zentralen Inhalte sowie historische Zusammenhänge sorgfältig zu erarbeiten. Übe, deine Texte klar zu gliedern und Fachbegriffe korrekt anzuwenden. Versuche außerdem, extreme und provozierende Formulierungen zu vermeiden und dich auf eine nüchterne, sachliche Analyse zu konzentrieren.

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

## Noten Copilot (Auszüge aus dem Feedback)

### Feedback

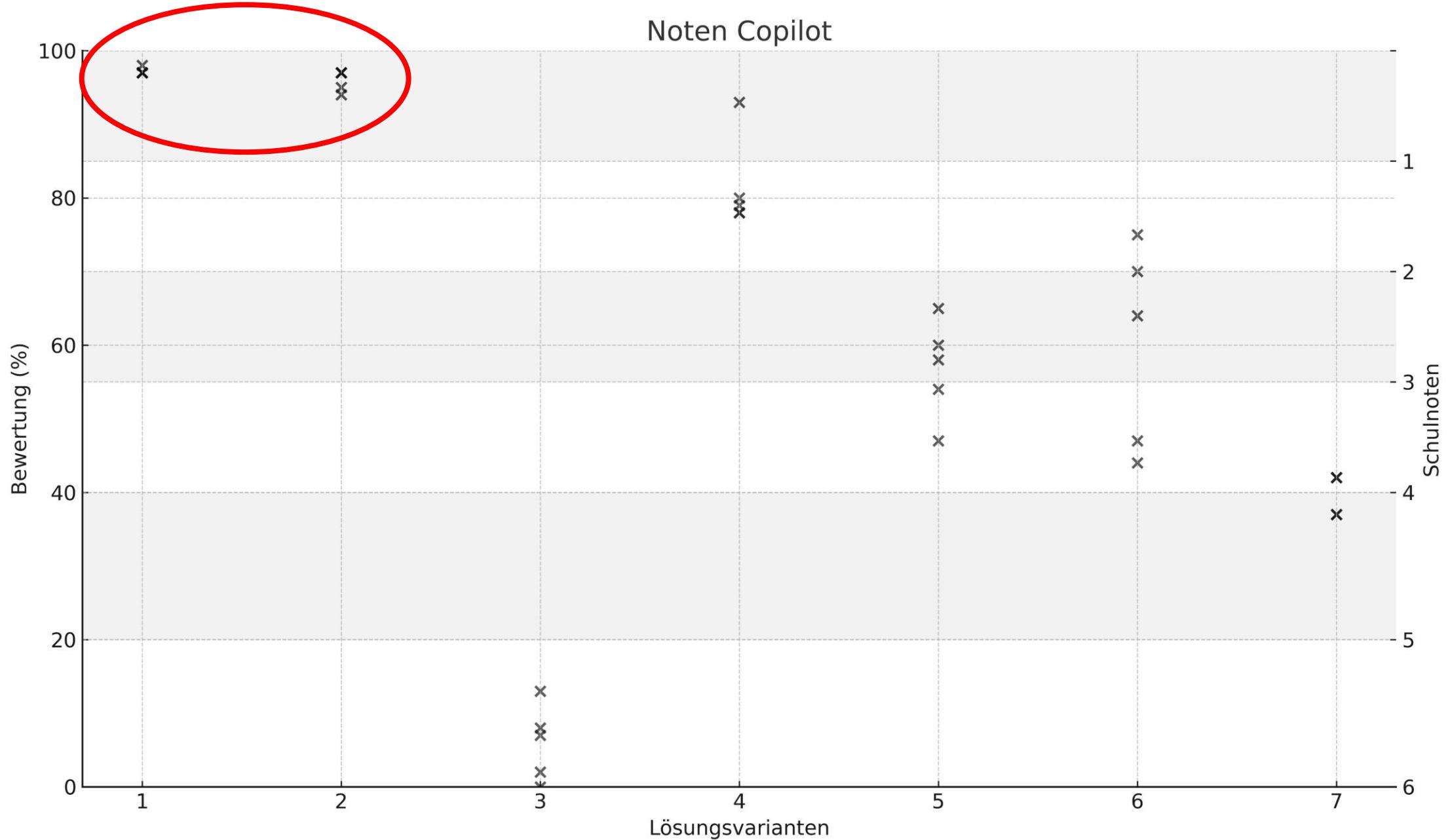
Deine Arbeit enthält gravierende inhaltliche und sprachliche Mängel. Besonders positiv hervorzuheben ist, dass Du versucht hast, mit direkten Quellenverweisen zu arbeiten. Leider weichen Deine inhaltlichen Ausführungen massiv von den tatsächlichen Aussagen der Quelle ab und enthalten zahlreiche unsachliche, historisch verzerrte und zudem rassistisch gefärbte Aussagen. Es ist wichtig, dass Du bei zukünftigen Arbeiten eine klare, sachliche und differenzierte Argumentation verfolgst. Arbeit gezielt daran, die historischen Fakten und die zugrunde liegenden ideologischen Zusammenhänge präzise zu erfassen und fachlich korrekt darzustellen. Als nächsten Schritt empfehle ich Dir, dir nochmal die zentrale Bedeutung der verwendeten historischen Quellen und deren Details (wie Verfasser, Titel, Erscheinungsjahr und Intention) anzusehen sowie den methodisch-kritischen Umgang mit historischen Quellen zu üben. Bitte vermeide in zukünftigen Arbeiten unsachliche und diskriminierende Inhalte, da diese den wissenschaftlichen Diskurs nicht fördern.

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

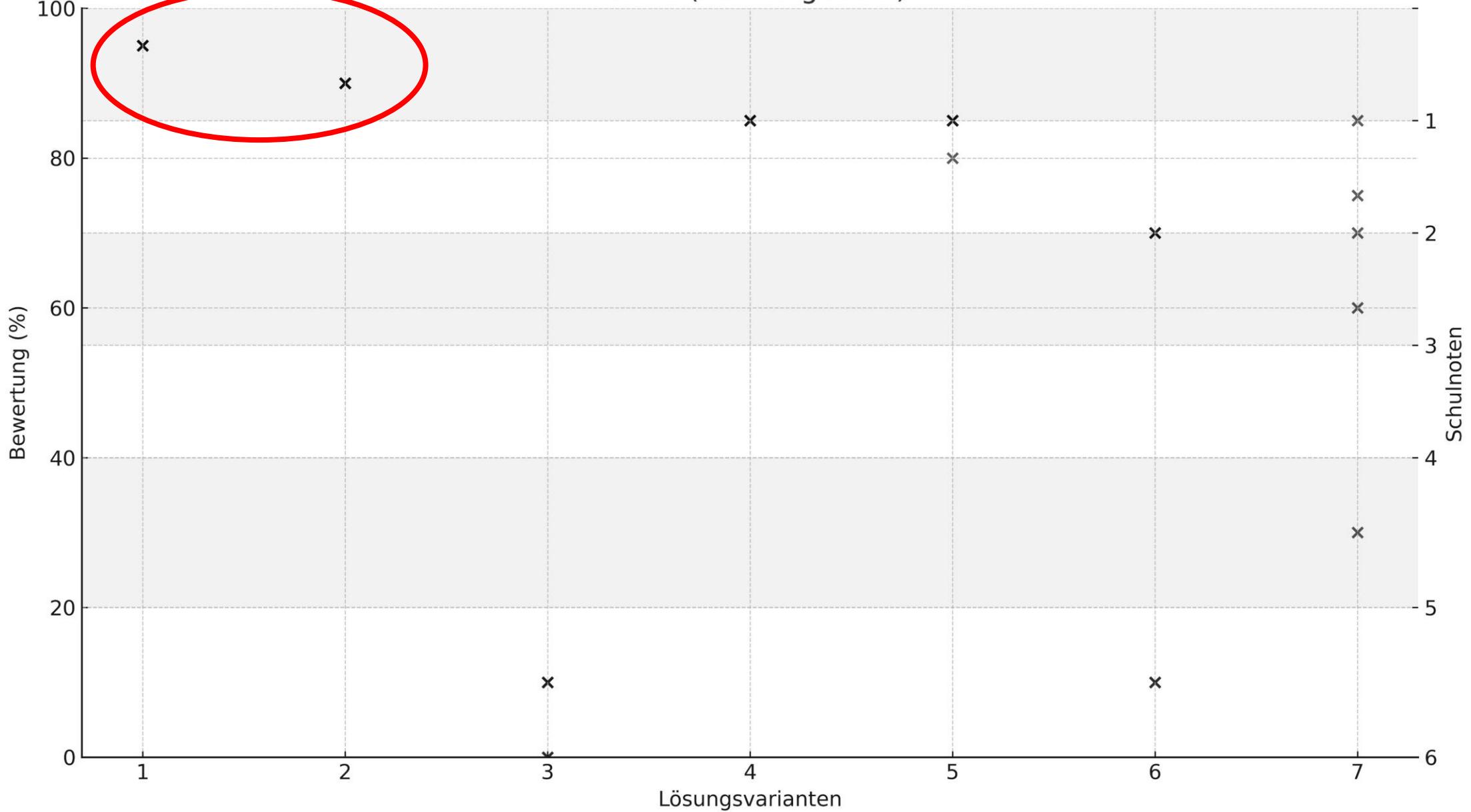
## Unzuverlässige Erkennung von **inhaltlichen Fehlern**

- kein „Internationaler Gerichtshof in Bagdad“, sondern in Den Haag
- Verwechslung von Kommunismus und Nationalsozialismus
- kein „Friedenspreis des Deutschen Börsenvereins“ für Charles Darwin 1939
- teils Fehlerberichte nur sehr allgemein und wenig konkret, ohne präzise Korrekturhinweise



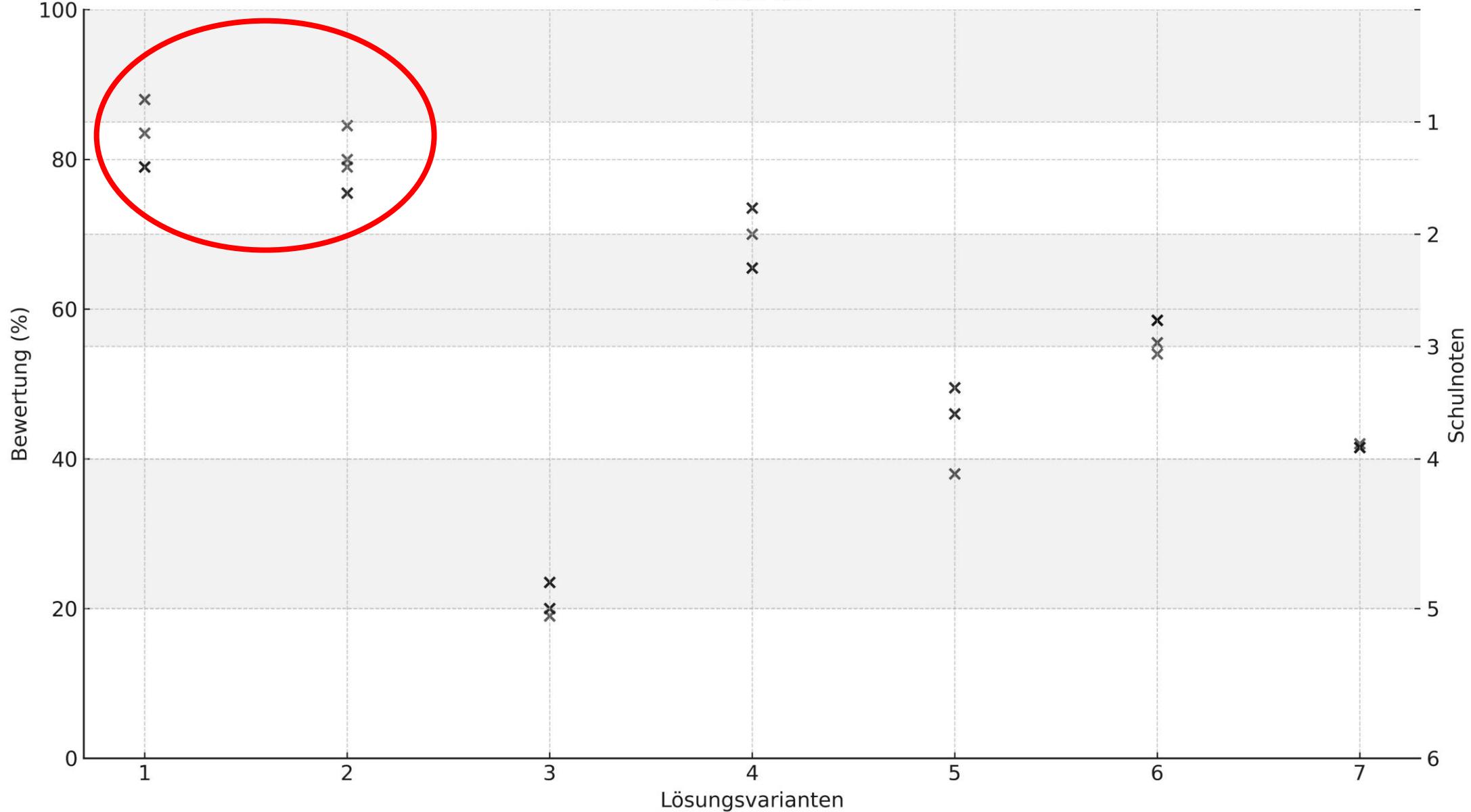
Quelle: Burkhardt / Erichsen / Klein (2025): Verteilung der Bewertungen der Lösungsvarianten bei Notencopilot bei jeweils 5 Durchläufen pro Lösung (Gesamtbewertung bei allen Teilaufgaben).

### Fobizz (nur Aufgabe 2)



Quelle: Burkhardt / Erichsen / Klein (2025): Verteilung der Bewertungen der Lösungsvarianten von Fobizz bei jeweils 5 Durchläufen pro Lösung (nur inhaltliches Kriterium bei Aufgabe 2).

# FelloFish



# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

## Inkonsistente Überführung von **Bewertungskriterien**

- FelloFish: Die im Erwartungshorizont benannten Kriterien können direkt einzeln ins System übernommen werden.
- Fobizz verwendet drei Standardkriterien, die verändert, gelöscht oder ergänzt werden können, erzeugt aber teils abweichende Bewertungskriterien bei der Bewertung.
- Noten Copilot generiert auf Grundlage der Aufgabe eigene Bewertungskriterien, auch wenn ein Erwartungshorizont vorliegt, und vergibt maximale Punkte für jedes Kriterium. Dabei können bei gleichen Vorgaben unterschiedliche Kriterien entstehen.

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

## Oberflächliches Feedback

- unkonkret, wenig spezifisch
- kaum vertiefte Feedbackformen, die sich auf inhaltliche Konzepte, präzise Fehlererklärungen, konkrete Verbesserungsschritte und die Förderung von metakognitiven Fähigkeiten beziehen
- Feedback oft sehr allgemein und markiert keine konkreten Fehlerstellen oder inhaltlichen Schwächen präzise
- z. B. kaum konkrete Vorschläge zur Verbesserung historischer Quellenkritik oder Urteilsbildung

# KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht? | Problemfelder

---

1. Hohe **Volatilität** in den Bewertungen
2. Unzureichende Erkennung und Missbilligung von **verfassungsfeindlichen Inhalten**
3. Unzuverlässige Erkennung von **inhaltlichen Fehlern**
4. Zu gute Bewertung von **KI-generierte Lösungen**
5. Inkonsistente Überführung von **Bewertungskriterien**
6. **Oberflächliches Feedback**

## IMPULSE FÜR DEN GESCHICHTSUNTERRICHT

■ Hannes Burkhardt, Jakob Erichsen, Malte Klein

### KI als Unterstützung für Schülerfeedback und Notengebung im Geschichtsunterricht?

Ein Problemaufriss

#### 1. Einleitung

Gegenwärtig erleben wir eine Konjunktur kommerzieller Angebote, die sich an Schulen und Lehrkräfte richten und verschiedenste KI-Anwendungen für den schulischen Alltag bereitstellen. Teil dieser Produktpalette sind digitale Werkzeuge, die eine KI-basierte Korrektur und ein KI-basiertes Schülerfeedback als Unterstützung für den Unterricht und ggf. für die Notenfindung anbieten. Beispiele sind der aktuelle Marktführer Fobizz, die Software FelloFish (ehemals Fiete.ai) oder die Anwendung Noten Copilot, die bereits im Namen explizit eine Unterstützung bei der Notenfindung verspricht. Auch die Anwendungen KorrekturKumpel, Cornelius oder Cornelsen.AI versprechen KI-basierte Unterstützung bei der Korrektur von Schülerarbeiten.

Die Anbieter werben teils mit einer Zeitersparnis bei der Korrektur: „Korrigiere Klausuren, Tests und Hausaufgaben bis zu 70 % schneller“ verspricht Noten Copilot, während KorrekturKumpel offensiv mit dem Claim „Rotstift weg – Freizeit her“ wirbt (beides Stand Mai 2025). Die Anbieter zielen mit diesen Versprechen auf die bei Lehrkräften allgemein wenig beliebte Tätigkeit des Korrigierens und adressieren strukturelle Probleme des Bildungssystems wie die hohe Arbeitsbelastung, die auch bei Geschichtslehrkräften aus dem Korrekturaufwand resultiert (Anders 2023). Vor diesem Hintergrund erscheint der Einsatz von KI bei der Korrektur zur Zeitersparnis verlockend, auch für Gewerkschaften: Die GEW stellt beispielsweise ihren Mitgliedern den Noten Copilot teils vergünstigt zur Verfügung (Stand Mai 2025).<sup>1</sup>

Burkhardt, Hannes / Erichsen, Jakob / Klein, Malte:  
Künstliche Intelligenz als  
Unterstützung für  
Schülerfeedback und  
Notengebung im  
Geschichtsunterricht? Ein  
Problemaufriss.

In: *geschichte für heute* 18  
(2025). Heft 4. S. 49–69.

# Bias im Notenspiegel?

Problemfelder der Nutzung Künstlicher Intelligenz  
als Unterstützung für Notengebung und Feedback

*KI-Werkstattgespräche zur Mittagspause an der EUF  
am 17. November 2025*

**Dr. Hannes Burkhardt**  
OStR, Educational Engineer

Landesprogramm „Zukunft Schule im digitalen Zeitalter“

Seminar für Geschichte und Geschichtsdidaktik  
Europa-Universität Flensburg (EUF)

