KLEIN, PASCAL; DAHLKEMPER, MERTEN N.; LAHME, SIMON Z.

Spot the Bot: Wahrnehmung von ChatGPT-Antworten auf Physikfragen

Basierend auf dem PRPER-Artikel (Mai 2023)
10.1103/PhysRevPhysEducRes.19.010142

Link zum Paper

Motivation

- Anekdotische Beobachtung: ChatGPT wird von Studierenden für Physikaufgaben genutzt
- Auffallend: Antworten auf Physikaufgaben klingen überzeugend – sind aber häufig fehlerhaft Antwort (Kortemeyer 2023, West 2023, Gregorcic & Pendrill 2023)
- Gefahr der unreflektierten ChatGPT-Nutzung
  (s. auch Verstehensillusion, Kulgemeyer & Wittwer 2023)
- (Feb 23): keine empirische Erkenntnisse zur Rezeption von ChatGPT-Antworten in Physik durch Lernende

Forschungsfragen

1. Wie bewerten Physikstudierende ChatGPT-Antworten zu Mechanik-Verstehensaufgaben in Hinblick auf ihre *“scientific accuracy”* sowie ihre *“linguistic quality”* verglichen mit einer maskierten Musterlösung?
2. Welchen Einfluss hat die (selbsteingeschätzte) Expertise auf diese Bewertung?
3. Existieren potentielle Unterschiede bezüglich der “*scientific accuracy*” weiterhin, wenn die *“linguistic quality”* berücksichtigt wird?

Design

Präsentation der Fragen:

- 3 Physikaufgaben unterschiedlicher Schwierigkeitsgrade (Rollbewegung, Wellenausbreitung, Strömungsmechanik)
- Je Aufgabe 4 Antworten (3x ChatGPT 3.5, 1x maskierte Musterlösung der Autoren) in radomisierter Reihenfolge

Validierung und Pilotierung:

Aufgaben und Antworten von Expert:innen (Physik-Fakultät) validiert und mit Studierenden aus der AG pilotiert

Aufgabe 1 - Aufgabenstellung

Beim Vorlesungsexperiment "Wettrennen der Zylinder am Hang" rollten ein Vollzylinder (Radius R, Länge L, Masse M) und ein Hohlzylinder (Außenradius R, Länge L, Masse M) eine schiefe Ebene hinab. Zu Beginn waren beide Zylinder in Ruhe und befanden sich auf gleicher Ausgangshöhe. Welcher Zylinder erreichte als erstes das untere Ende der Ebene? Begründen Sie Ihre Antwort!

*** Ohne die Aufgabe selbst zu lösen: Schätzen Sie ein, wie viele Punkte Sie in einer Klausur bei dieser Aufgabe erhalten hätten (0-6 Punkte)?**

Bitte wählen Sie eine der folgenden Antworten:

0 | 1 | 2 | 3 | 4 | 5 | 6

Datenerhebung

Antwort: **Der Hohlzylinder wird als erstes das untere Ende der Ebene erreichen, da er ein niedrigeres Trägheitsmoment hat, was bedeutet, dass er weniger Widerstand gegen die Rotationsbewegung hat. Dies macht es dem Hohlzylinder einfacher, die Schräge hinabzurollen und erreicht daher schneller als der Vollzylinder das untere Ende.**

✱ **Bewerten Sie die Antwort bitte anhand folgender Kriterien:**

	sehr niedrig	niedrig	mittel	hoch	sehr hoch
Fachliche Korrektheit	○	○	○	○	○
Grad der Vollständigkeit	○	○	○	○	○
Verständlichkeit	○	○	○	○	○
Sprachliche Qualität	○	○	○	○	○
Eignung als Musterlösung	○	○	○	○	○

Beschreibung der Stichprobe

Insgesamt 94 Studierende aus zwei Vorlesungen (Experimentalphysik I + III)

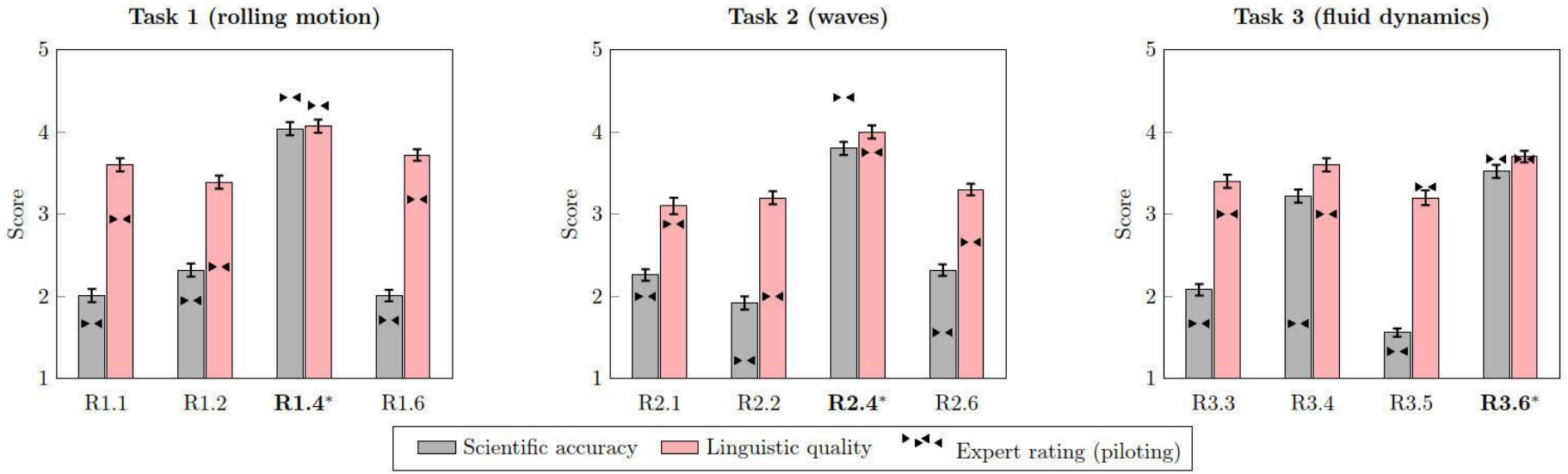
Demographie

76% Mono-Bachelor, 16% Lehramt
77% Erstsemester, 21% Drittsemester
21% weiblich, 73% männlich, 1% divers

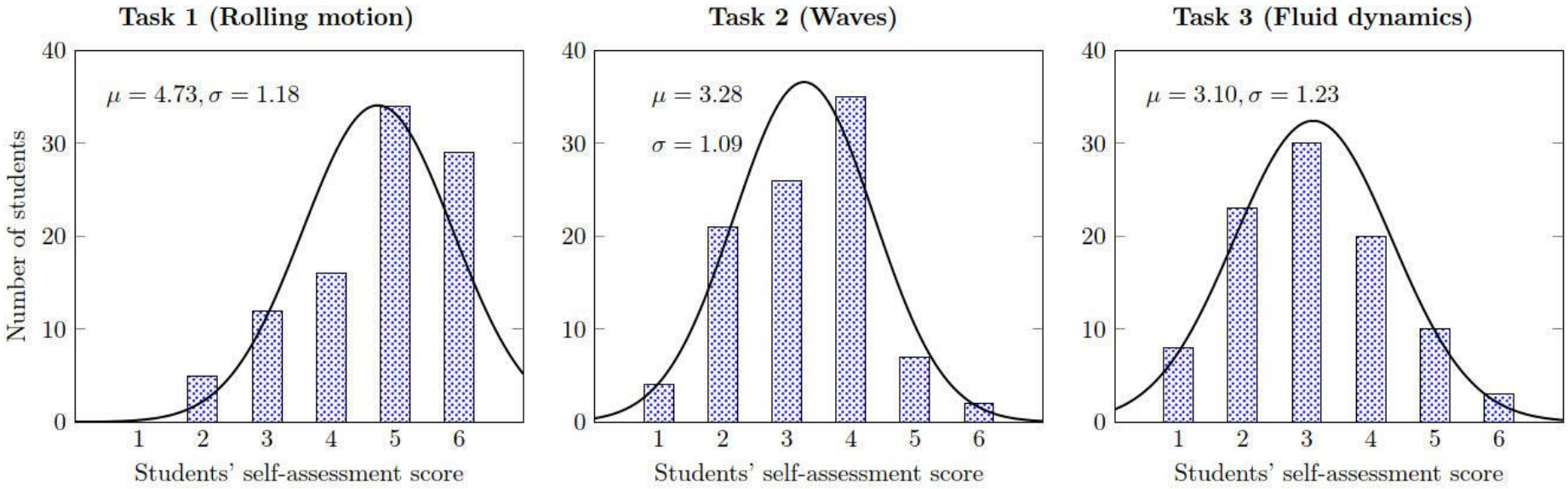
Bisherige Nutzung von ChatGPT (02/23)

84% haben von ChatGPT gehört, 48% schon mal genutzt
26% schon mal im Physik-Kontext genutzt

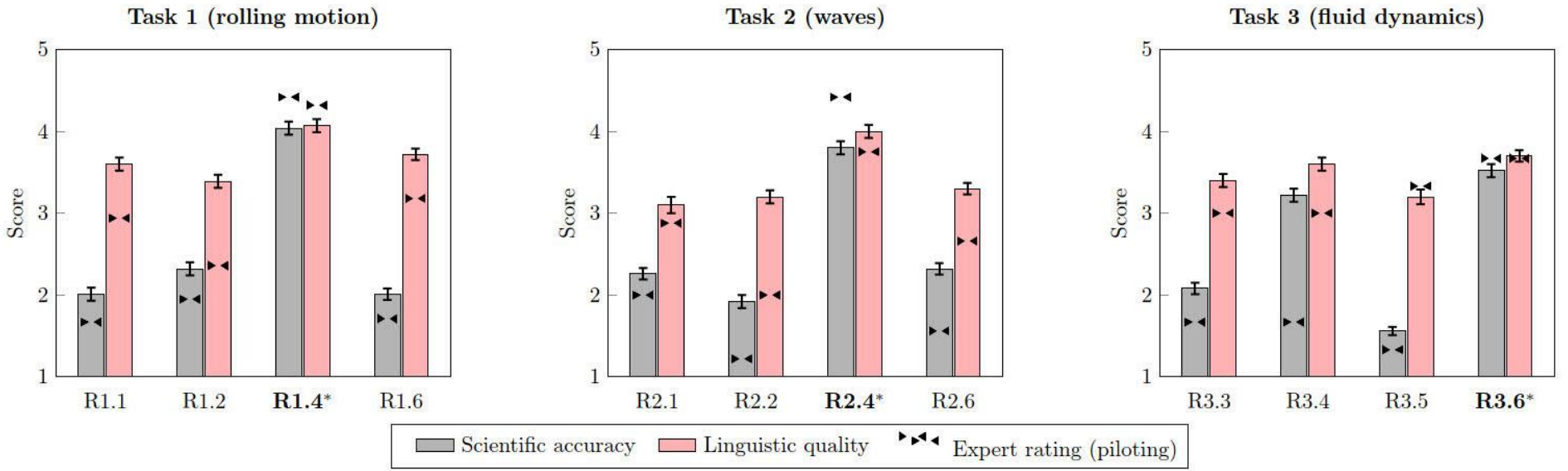
FF1: Einschätzung der Antworten



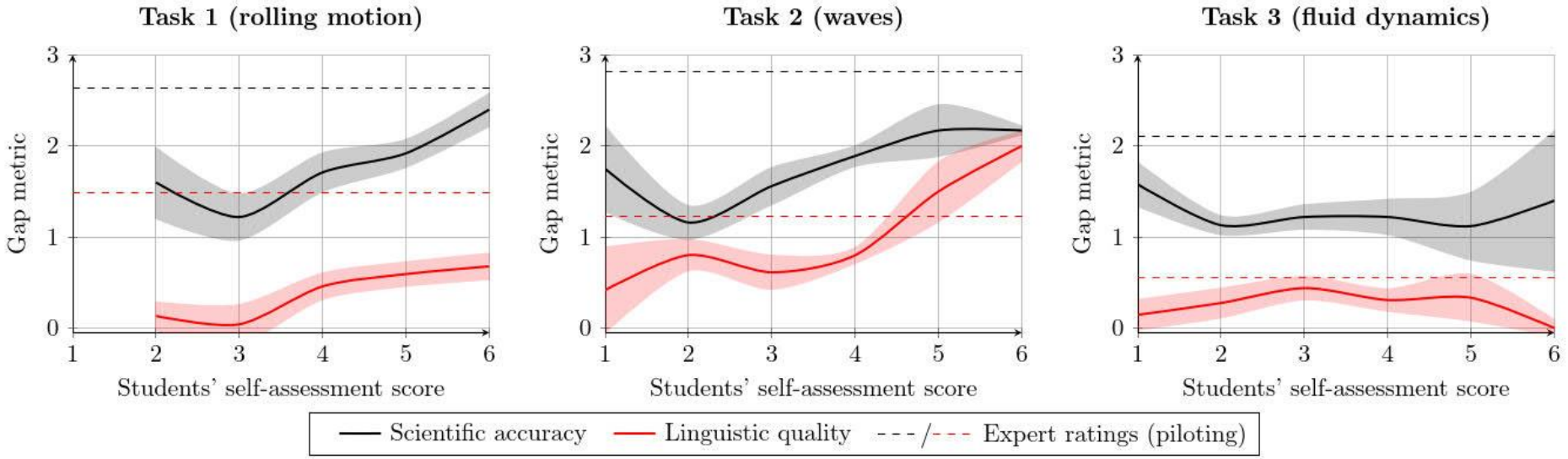
FF2: Selbsteingeschätzte Expertise



FF1: Einschätzung der Antworten



FF2: Abhängigkeit von der Expertise



"Gap Metric": MW-Differenz zw. Einschätzung der ChatGPT-Antworten und der Musterlösung

FF3: Einfluss der sprachlichen Qualität

Model 2: Einfluss der dargebotenen Antworten auf die „scientific accuracy“
Model 1: zusätzlich Korrektur auf die „linguistic quality“ (Kovariate)

	η^2	
	Model 1 (ANCOVA)	Model 2 (ANOVA)
Question 1 (rolling motion)	0.46	0.53
Question 2 (waves)	0.26	0.49
Question 3 (fluid dynamics)	0.43	0.57

Diskussion und Fazit

- Je schwieriger eine Aufgabe für die Studierenden ist, desto empfänglicher scheinen sie für eine plausible, aber falsche/lückenhafte ChatGPT-Antwort
- Je mehr sie über ein Thema wissen, desto akkurater können sie die Antwort einschätzen
- Falsche Antworten erscheinen auch deshalb glaubwürdig, weil die Art der sprachlichen Darbietung sehr gelungen ist.

Zukünftige Forschungrichtungen

- Welchen Einfluss hat das (vermeintliche) Wissen über den Urheber einer Antwort auf die Einschätzungen?
- Welche Fehlkonzepte können in ChatGPT-Antworten gefunden werden?
- Unterscheiden sich die Ergebnisse mit Antworten von ChatGPT-4 und zukünftigen LLM-Tools?

Link zum Paper

Referenzen

- B. Gregorcic and A.-M. Pendrill, ChatGPT and the frustrated Socrates, *Phys. Educ*. 58, 035021 (2023).
- C. Kulgemeyer and J. Wittwer, Misconceptions in physics explainer videos and the illusion of understanding: An experimental study, *Int. J. Sci. Math. Educ.* 21, 417 (2023).
- G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Phys. Rev. Phys. Educ. Res.* 19, 010132 (2023).
- C. G. West, *AI and the FCI: Can ChatGPT project an understanding of introductory physics?*, arXiv:2303.01067.