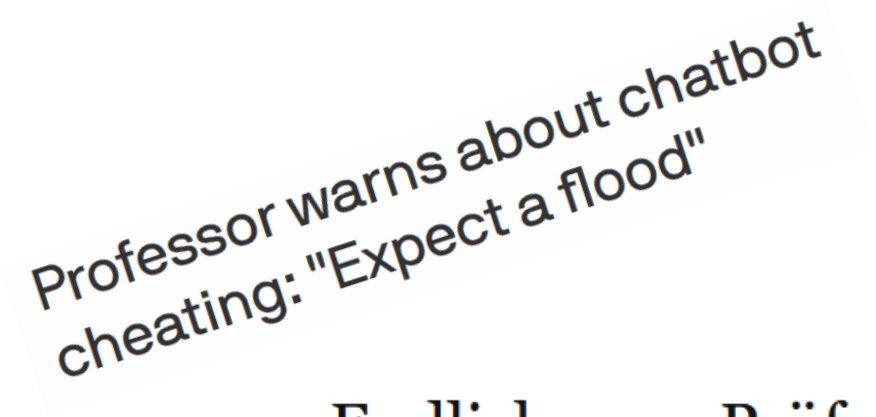
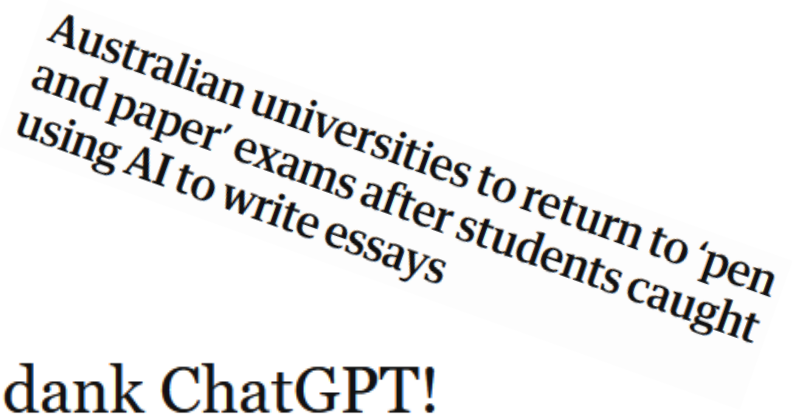
Universität Potsdam

Anfälligkeit von KI-Generatoren hinsichtlich E-Assessments

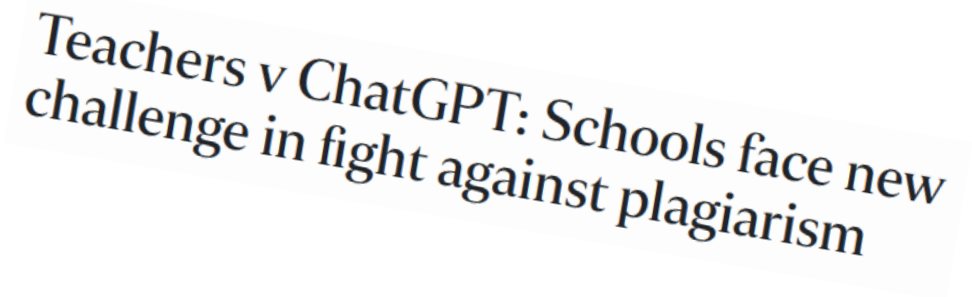
Dr. Tobias Moebert und Dr. Evgenia Samoilova

KI und Lehre – Der Untergang des Abendlandes?





Endlich neue Prüfungen dank ChatGPT!



Problemstellung: Auswirkungen von generativen KI-Modellen auf die akademische Integrität

- In einer Welt mit leicht zugänglichen LLM-Modellen, was bedeutet KI für das, **was** und **wie** wir es in der Hochschulbildung bewerten?

- Welche kurzfristigen Maßnahmen können/sollten wir ergreifen?

- Herausforderungen und **Chancen**: Welche Risiken bestehen bei der Verwendung von LLM in Prüfungssituationen und welche **Vorteile** könnten sie bieten?

Studie: Forschungsfragen

Wie schneiden Studierende ohne Fachkenntnisse und LLM-Erfahrung, die ChatGPT 3.5 nutzen, in Prüfungen im Vergleich zu Studierenden des Kurses ab?

Explorative Untersuchung:

- Technische Aspekte der Fragen
- Andere Faktoren: z. B. faktisches Wissen im Vergleich zu kreativen Aufgaben
- Interaktion der Studierenden mit ChatGPT im Laufe der Zeit (5 Monate)

Studienaufbau (I)
nicht-probabilistische Stichprobe der Freiwilligen

Fächer	Kurse	Klausuren	Anzahl der Fragen	Anzahl der Studierenden (Verteilung der vorherigen Prüfungsergebnisse)
Erziehungswissenschaften	2	9	401	969
Psychologie	2	4	210	700
BWL	2	2	107	384
Sozialwissenschaften	2	2	56	203
Informatik	1	1	40	57
Agrarökologie	1	2	49	31
Linguistik	1	1	17	16
Mathematik	1	1	6	65
	12	**22**	**886**	**2.425**

22 Moodle- und Papierklausuren (Sommer20-Winter23)

Studienaufbau (II) – Fachfremde Personen

- **5 WHKs** (Chemie, Geoökologie, Soziologie/VWL, Informatik)
- Zeitraum: Mai-September 2023
- Arbeitsvertrag + NDA
- **3 fachfremde** Personen pro Klausur
- **Kein Zeitlimit**

Studienaufbau (III) – standardisiertes Protokoll

- Anleitung für Durchführung (standardisiertes Protokoll) in 18 Schritten
- keine Interpretation von Fragen oder Antworten + kein Prompting (Übertragung möglichst 1:1): das **Basis-Szenario ohne individuelle Anpassungen**

Studienaufbau (IV) - Merkmale der Fragen

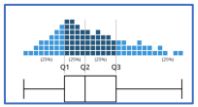
- zusätzliches Ausfüllen eines Begleitformulars
- 10 Punkte pro bearbeiteter Frage, z.B.
  - Prompt an ChatGPT
  - Antwort von ChatGPT
  - Übertragbarkeit 1:1 möglich (Frage / Antwort)
  - Details zur Vollständigkeit der Übertragung (Frage / Antwort)
  - Bilder enthalten
  - Strategie für die Übertragung
  - Schwierigkeiten bei der Übertragung
  - **Fragen-Features…**

Studienaufbau (V) – Fragentypen/Features

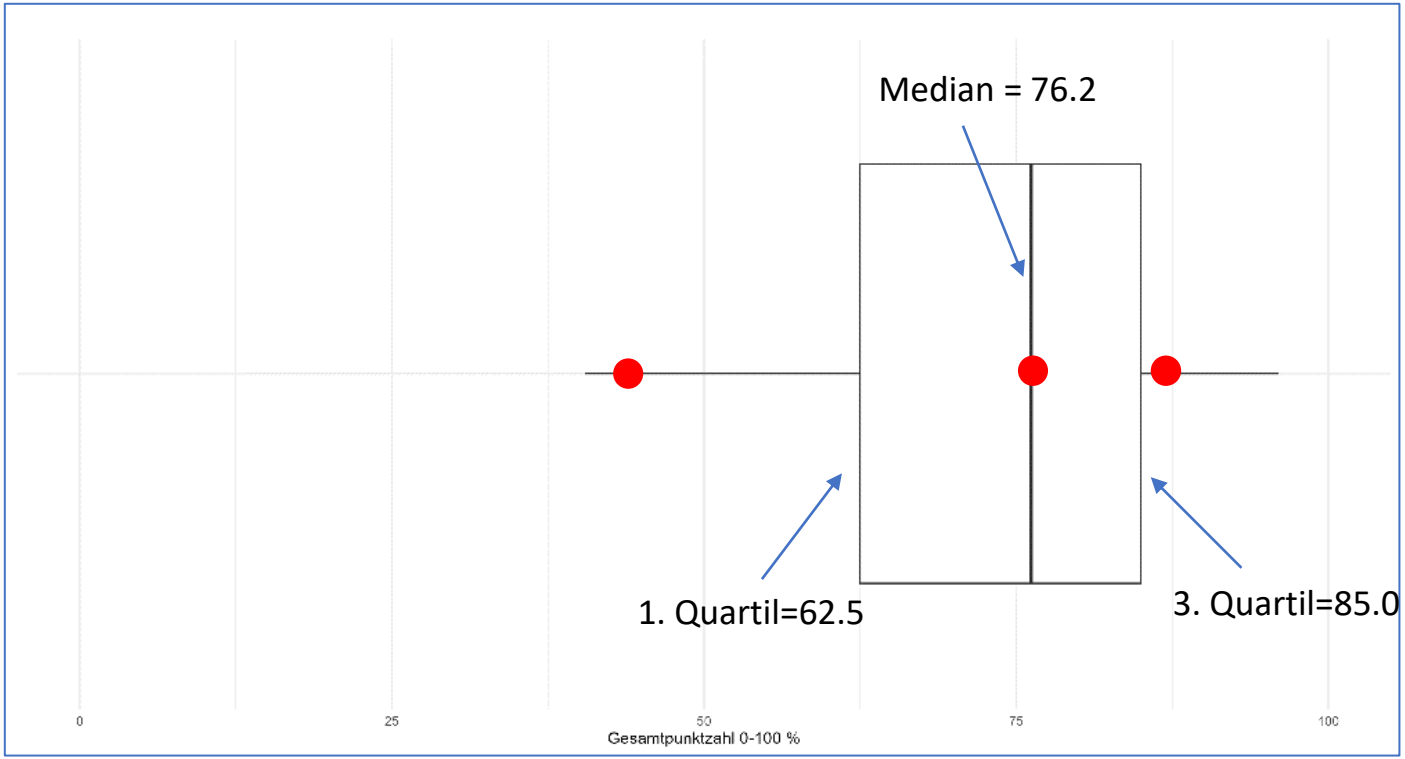
- zunächst Klassifizierung anhand der Moodle-Fragentyp (z.B. Freitext, Multiple-Choice, Lückentext, Wahr-oder-Falsch)
- unscharf und ungleichmäßige Verwendung innerhalb der Klausuren
- Klassifizierung anhand verwendeter **Features** (Mehrfachauswahl möglich)
    - **Radiobutton**
    - **Checkbox**
    - **Dropdown**
    - **Eingabefeld**
    - **WYSIWYG-Editor**
    - **Drag & Drop**

Studierende versus ChatGPT 3.5: Visualisierung der Ergebnisse

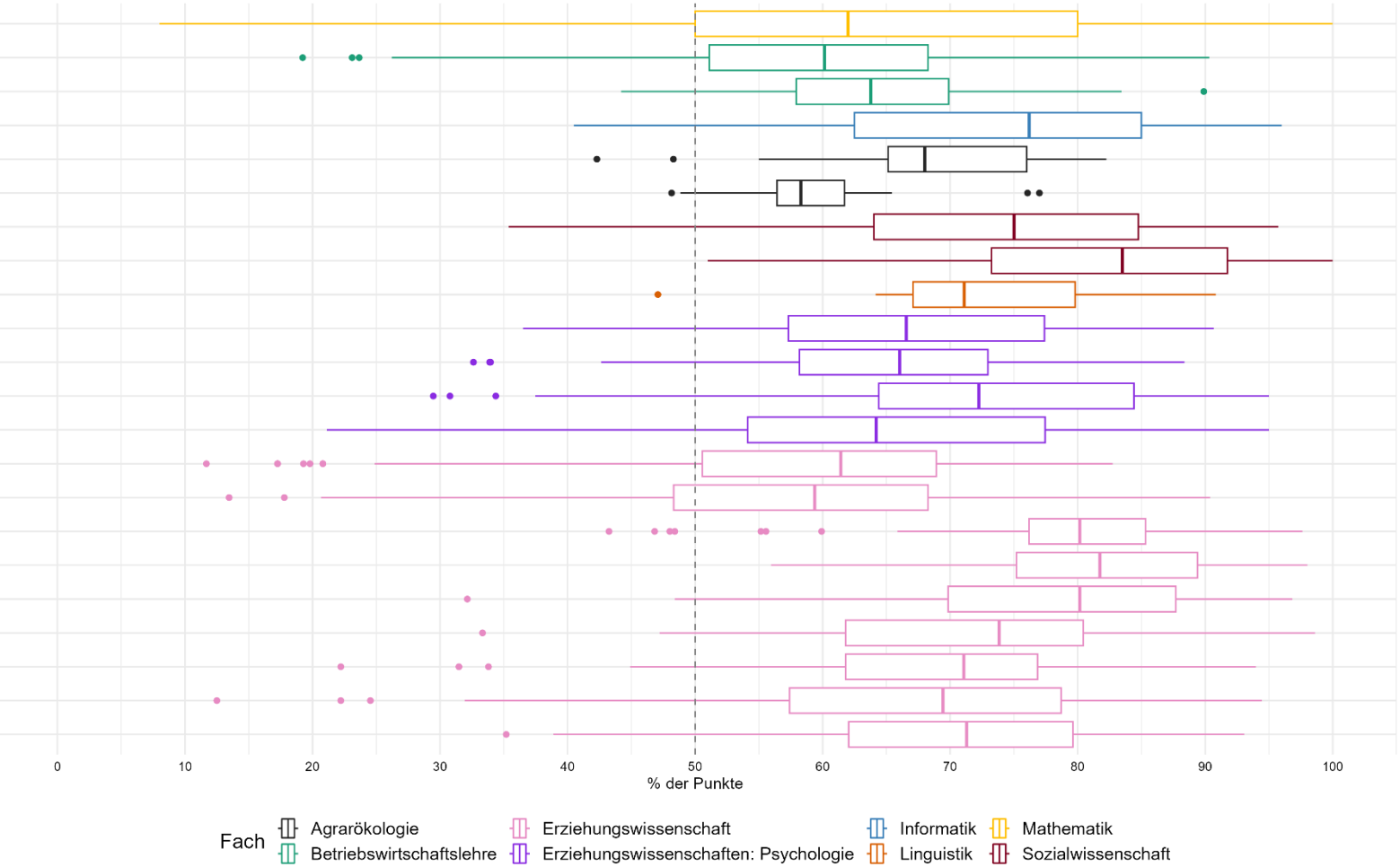
Studierende:



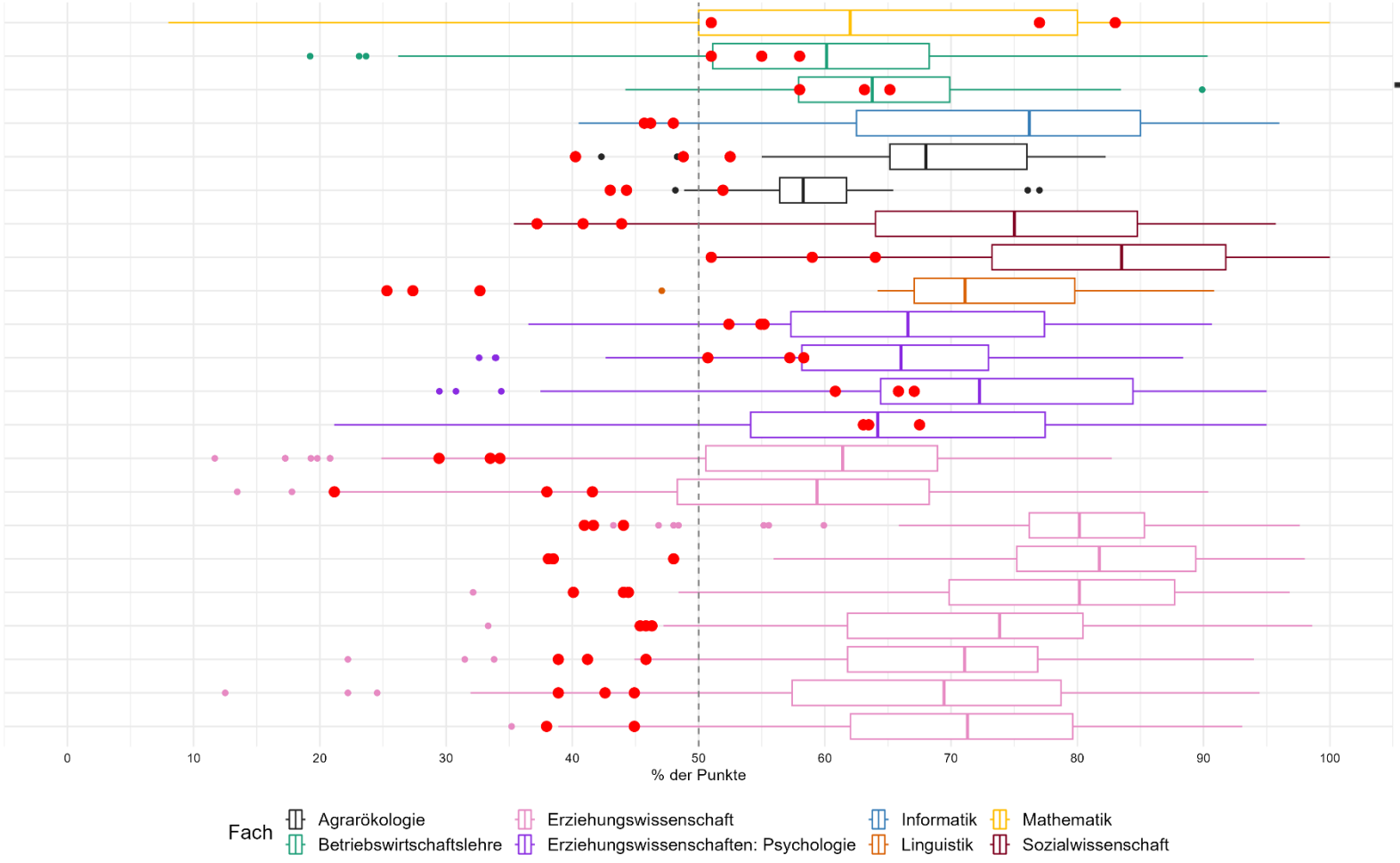
ChatGPT 3.5::



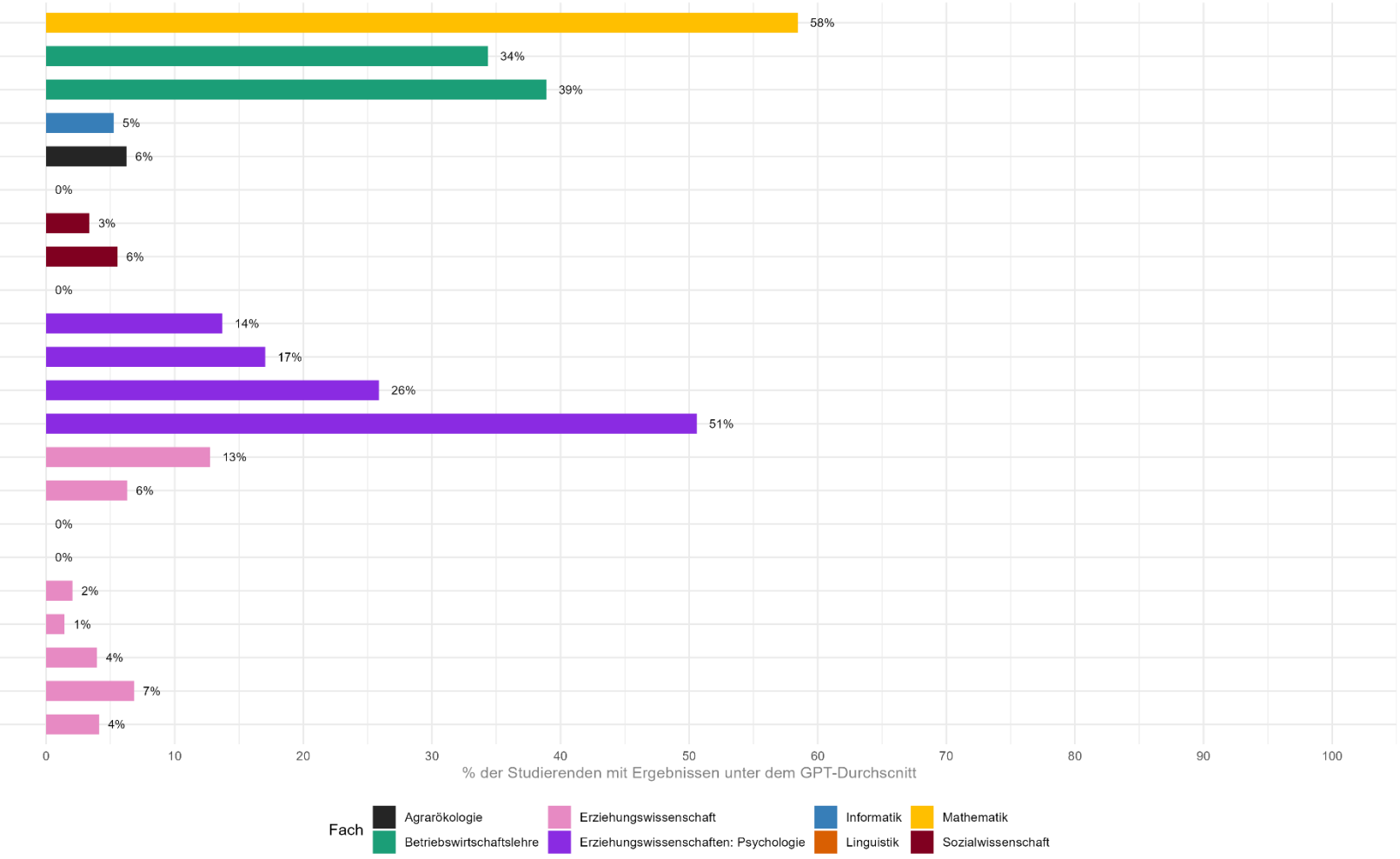
Verteilung der Gesamtpunktzahl der Prüfung (0-100 %), 22 Klausuren & 12 Kurse



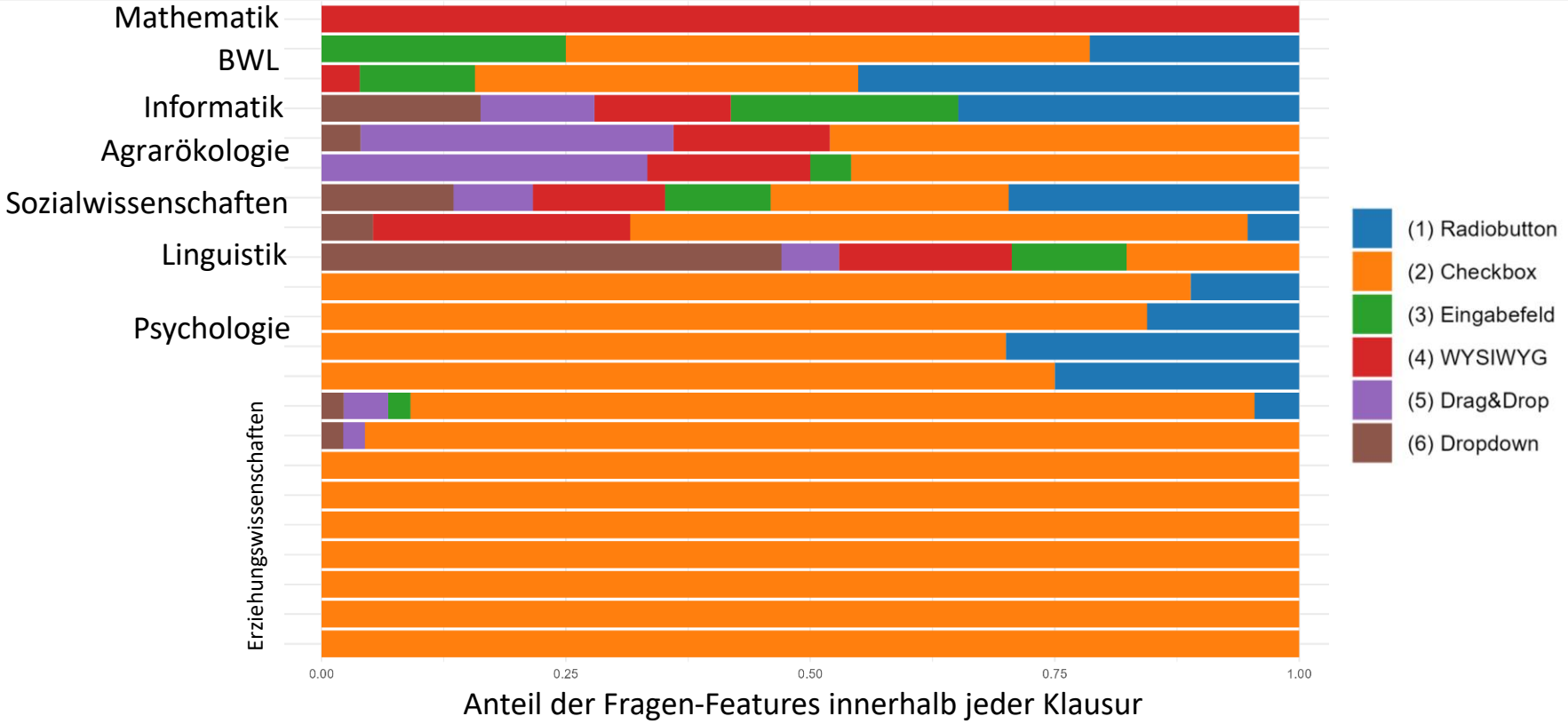
Verteilung der Gesamtpunktzahl der Prüfung (0-100 %), 22 Klausuren & 12 Kurse



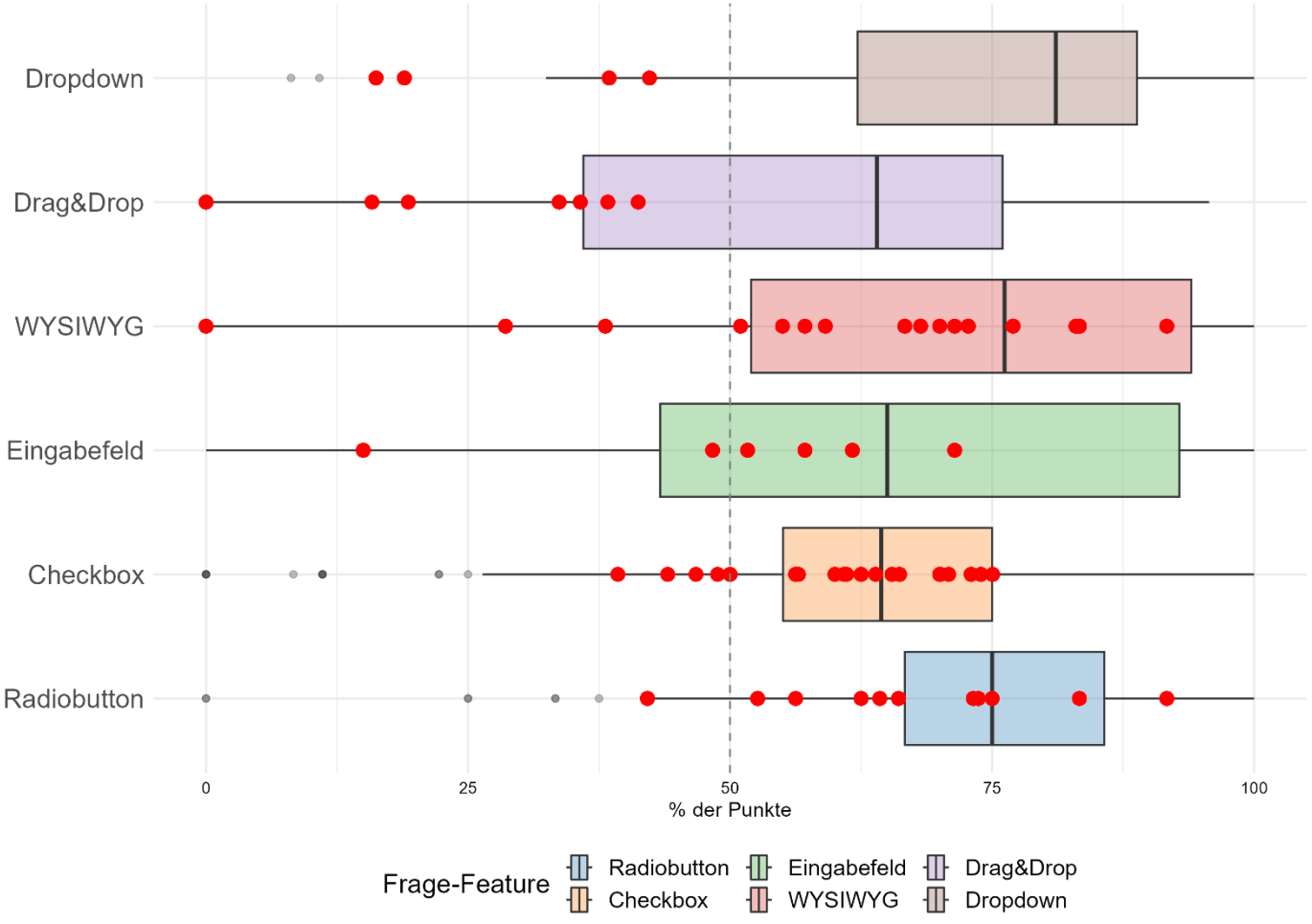
Verteilung der Gesamtpunktzahl der Prüfung (0-100 %), 22 Klausuren & 12 Kurse



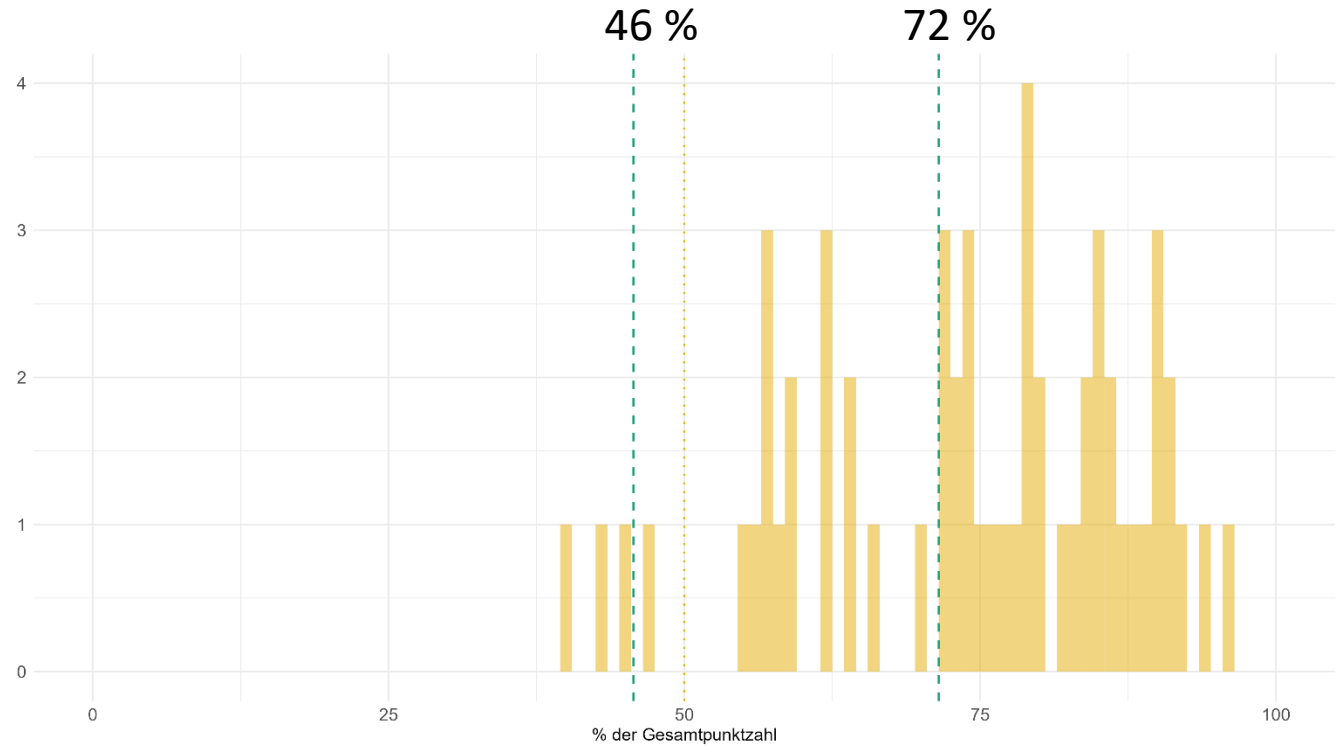
Frage-Features



Frage-Features:



Fallstudie Medientechnologie: Basis-Szenario + freies Bearbeiten



Kurs: Medientechnologie

Rückmeldung Lehrpersonen

- **Methodik**: einzelne Treffen (ca. 1 Stunde) mit den Lehrpersonen der 11 Kurse

- **Fokus**: Präsentation detaillierter Ergebnisse für einzelne Klausuren + Sammlung von Feedback und persönlichen Einsichten der Dozent*innen

- **Stimmung**: interessiert + überrascht (hinsichtlich der guten / schlechten Ergebnisse)

Rückmeldung Lehrpersonen – Probleme/Lösungsansatz

Problem	Vorgeschlagener Lösungsansatz
fehlendes Wissen über die Funktionsweise und den Nutzen von KI-Tools (Studierende / Lehrpersonen)	• Verfügbarkeit von Lizenzen • aktive Auseinandersetzung • universitäre Weiterbildungsangeboten
Unsicherheit über Kenntnisstand der Studierende zum Thema	• Universitätsweite Umfrage
Prüfung von Grundlagenwissen problematisch (leicht für LLMs; Präsenzprüfung logistisch schwierig)	• Abfrage des Grundlagenwissens in aktuelle Probleme einbetten
Sicherstellung der Vermittlung von Grundlagenwissen	? ? ?
Umgang mit Haus- / Abschlussarbeiten	• Grenzen der Anwendung von LLMs klar definieren • weniger Abfrage von Grundlagen → Fokus auf eigene Arbeit
Selbstständigkeitserklärung deckt LLMs möglicherweise nicht ausreichend ab	• Selbstständigkeitserklärung überarbeiten
Fehlende Unterstützung durch Führungsebene bei Problemen mit generativer KI in Lehre/Assessment	• Leitlinien/Workshops auf Universitätsebene

Zusammenfassung

- GPT 3.5 zeigt eine gemischte Performance in den Prüfungen.
- Höherer Schwierigkeitsgrad für Studierende ≠ Höherer Schwierigkeitsgrad für ChatGPT
- Ergebnisse mit GPT 3.5 stellen die konservativste Schätzung dar: wir erwarten, dass Studierende mit mehr Wissen, sowie fortschrittlicheren LLM-Tools besser abschneiden werden.
- Die an der Studie beteiligten Lehrenden zeigen sich besorgt/sind interessiert an die Thematik der generativen KI und fordern eine intensivere Auseinandersetzung.
- Bedürfnisse und Kompetenzen von Dozent*innen und Studierenden in Bezug auf generative KI benötigen weitere Untersuchung/Unterstützung.

“We must re-invent homework and develop teaching concepts that utilize these AI models in the same way as math utilizes the calculator: teach the general concepts first and then use AI tools to free up time for other learning objectives.”

Herbold, S., Hautli-Janisz, A., Heuer, U. et al. A large-scale comparison of human-written versus ChatGPT-generated essays. Sci Rep 13, 18617 (2023). https://doi.org/10.1038/s41598-023-45644-9

Dr. Tobias Moebert

Dr. Evgenia Samoilova

Universität Potsdam

Institut für Informatik und Computational Science

tobias.moebert@uni-potsdam.de

evgenia.samoilova@uni-potsdam.de