

WISSENSNUGGET #20

Prüfen im Zeitalter von ChatGPT – Wie kann ich meine Klausuren durch KI-resistente Aufgaben schützen?

York Hagmayer & Jacqueline Frank

Überblick

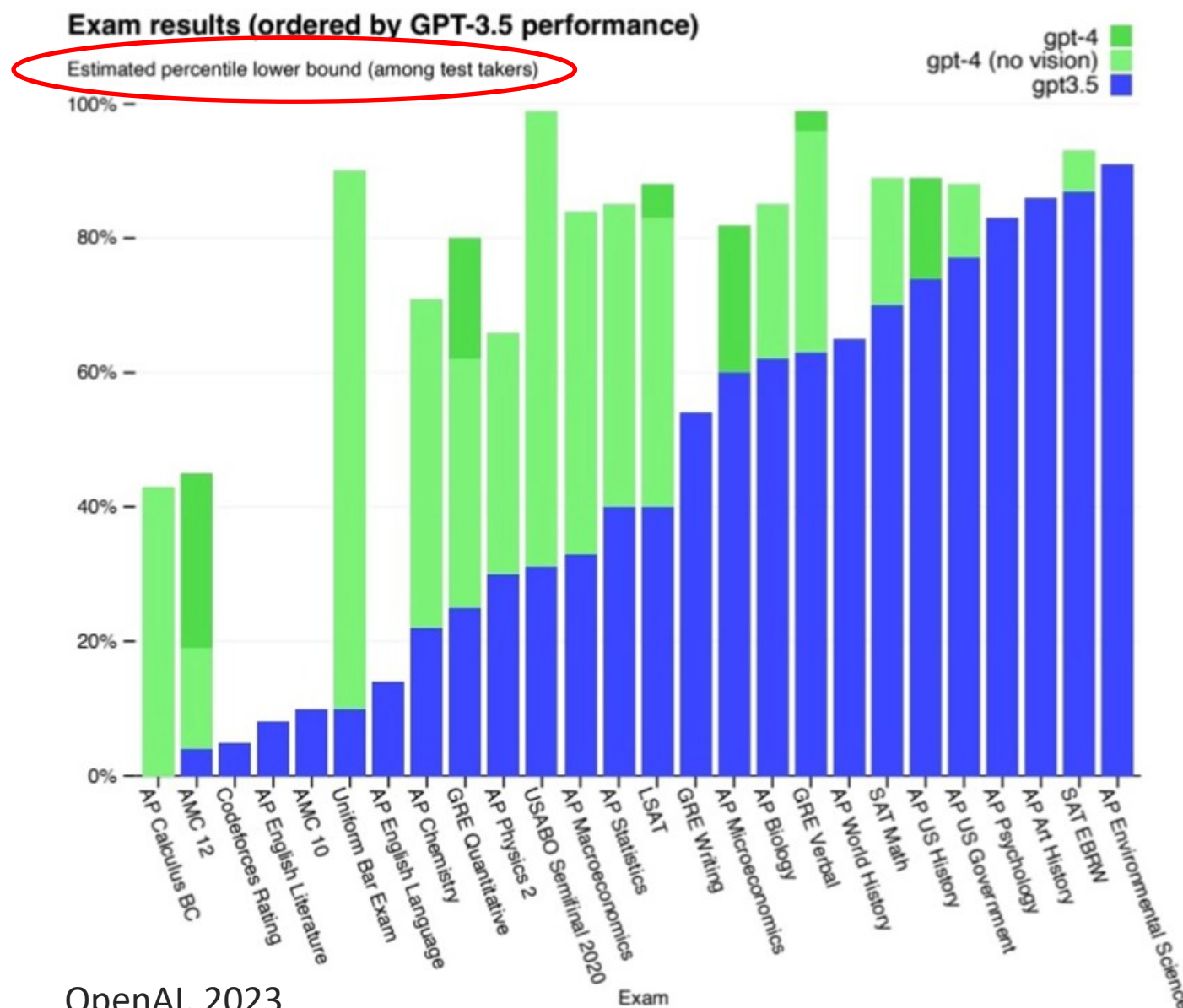
- Herausforderungen für Prüfende
- Strategien zum Umgang mit Herausforderungen:
KI-resistente Aufgaben als eine von vielen Ideen
- Konkrete Beispiele
- Studie 1: Test der Performanz von gängigen LLM
- Studie 2: Vergleich der Performanz Studierende und LLM
- Take Home und Ausblick

- Optional: Psychologie der KI-Nutzung

⇒ Fragen gerne zwischendurch im Chat

Herausforderungen –

(1) KI-Modelle sind extrem leistungsfähig



- Bereits ChatGPT 4 (3/2023) konnte sehr viele Prüfungen auf Universitätsniveau bestehen und sehr gute Noten erreichen (OpenAI, 2023)
- Akademische Schriften von KI sind von menschlichen Schriften nicht mehr unterscheidbar (Cotton et al., 2024)

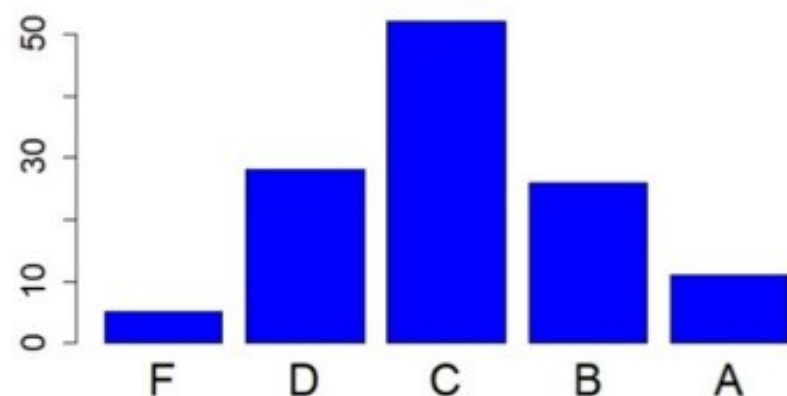
Herausforderungen –

(2) Studierende sind bereit zu täuschen

Fallstudie von Stanoyevitch (2024)

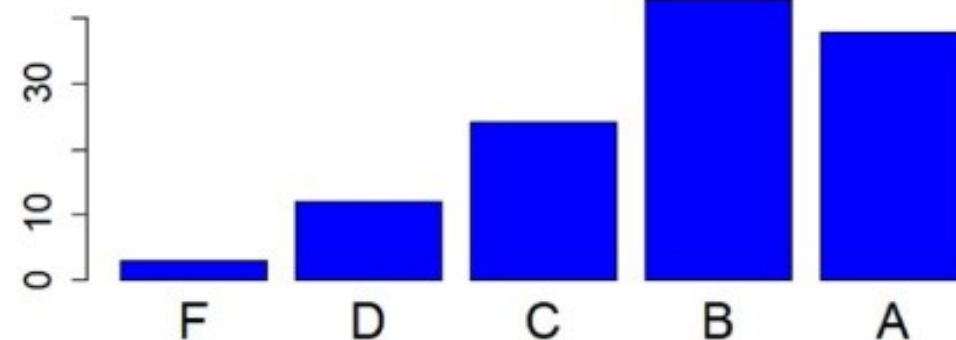
Langjährige Notenverteilung Präsenzklausur

Mat 131 Exam 1 Large Lecture Grades



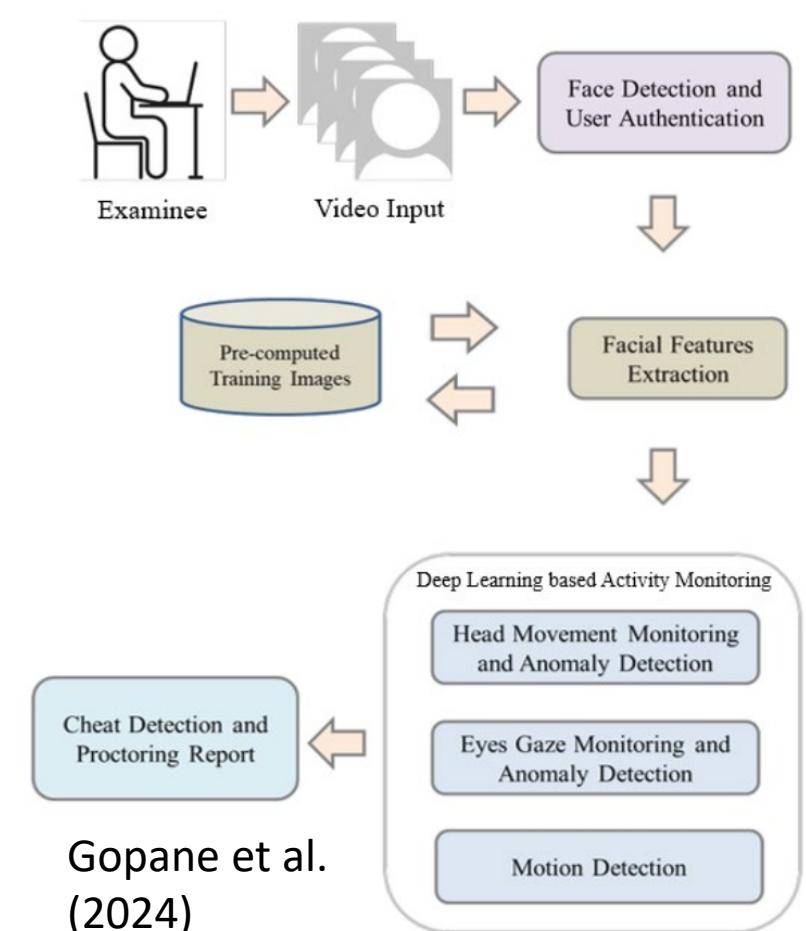
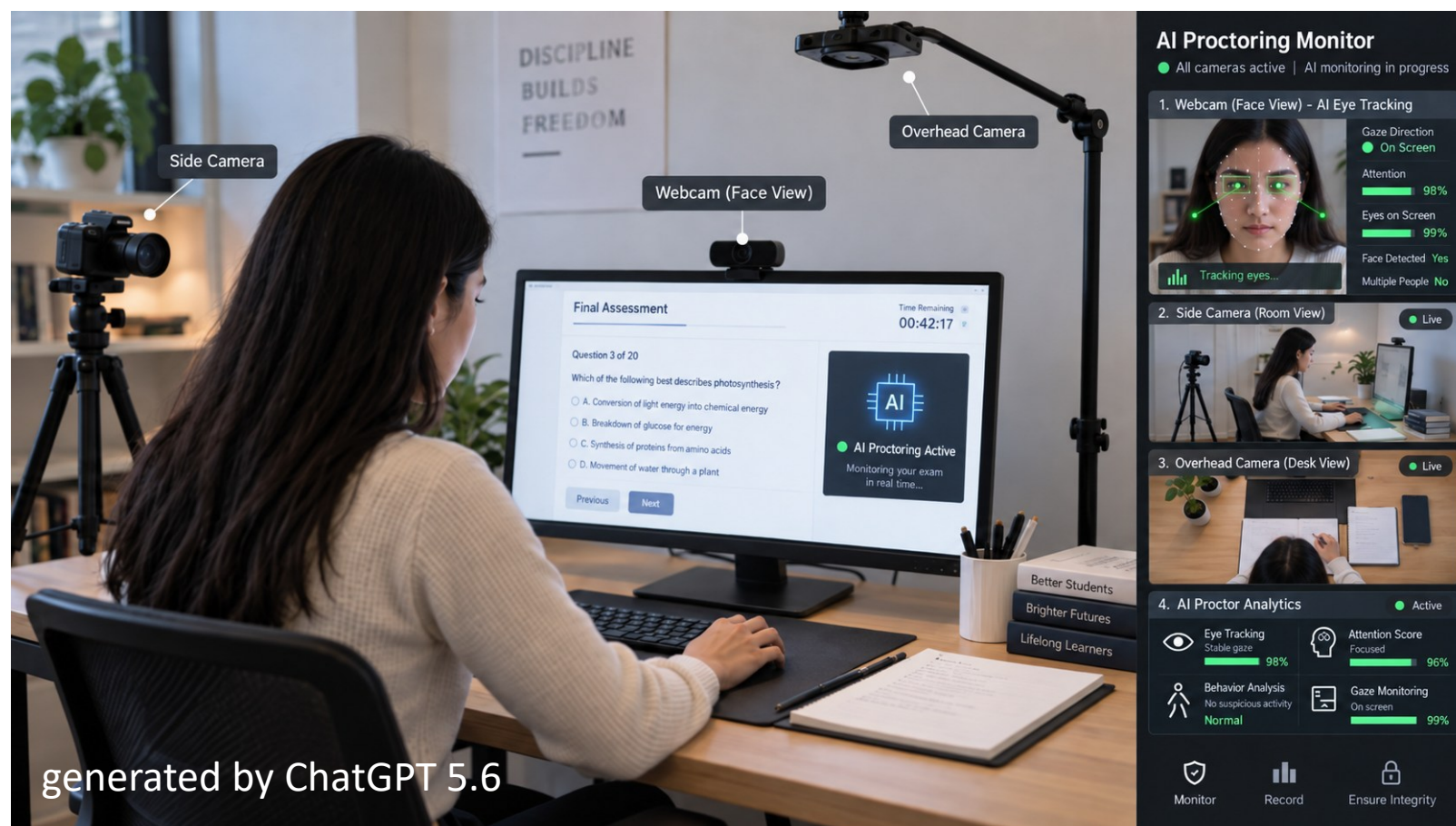
Notenverteilung Onlineklausur

Mat 131 Exam 3 Large Lecture Grades



- Studierende nutzen Hilfsmittel in Online-Assessments und verbessern so ihre Noten
- In Umfragen sagen bis zu 3 von 4 Studierenden, dass sie unerlaubte Hilfsmittel bei Online-Assessments verwenden (Susnjak & McIntosh, 2024; Jenkins, 2022)

Herausforderungen – (3) Die Verwendung unerlaubter generativer KI lässt sich im Einzelfall nur schwierig nachweisen



Herausforderungen – (4) Die zielgerichtete und kompetente Nutzung von KI ist eine für die Zukunft wichtige Kompetenz der Studierenden



Strategien zum Umgang mit den Herausforderungen

KI-Nutzung bei
Prüfungen
verhindern
(Proctoring)

⇒ Nur Wissen und Kompetenzen der Person sollen erfasst und bewertet werden

KI-Nutzung bei
bestimmten
Prüfungen oder
Aufgaben
erlauben

⇒ Je nach didaktischer Zielsetzung wird KI-Nutzung erlaubt oder unterbunden

KI-Nutzung
erlauben und
Aufgaben
anpassen

⇒ Lösung mit Hilfe von KI verlangt bestimmte, zu prüfende Kompetenzen

⇒ Aufgaben sind so gestaltet, dass KI per se schlecht in der Lage ist, diese perfekt zu lösen = KI-resistente Aufgaben

Schwächen von KI

Inhaltliches Verständnis fehlt

- Antwort wird auf Grundlage sprachlicher Muster generiert = assoziatives Vorgehen
- Es findet nicht automatisch eine Vorauswahl von relevanten Inhalten statt

Prompt-Sensitivität

- Kleine Änderungen können deutlich unterschiedliche Antworten bedingen
- Nicht optimale Prompts können zu suboptimalen Antworten führen
- Prompts werden „vergessen“

Probleme mit Quellen

- Referenzen werden z.T. erfunden
- Quellen werden z.T. falsch wiedergegeben
- Wenn keine Antwort gefunden wird, wird halluziniert und dies als Antwort gegeben

Visuelle Limitationen

- Wahrnehmung von Farben und Labels kann unzuverlässig sein

Positionseffekte bei Single und Multiple Choice

- Erste plausible Antwort wird bevorzugt

Fehlerkumulation bei mehrschrittigen Aufgaben

- Sinnvolles mehrstufiges Vorgehen muss expliziert werden

Biases in Trainingsdaten finden sich in Antworten

Disclaimer: Die KI Modelle werden immer besser und Entwickler arbeiten daran, diese Schwächen abzustellen

Beispiel aus <https://research.aimultiple.com/ai-hallucination/>

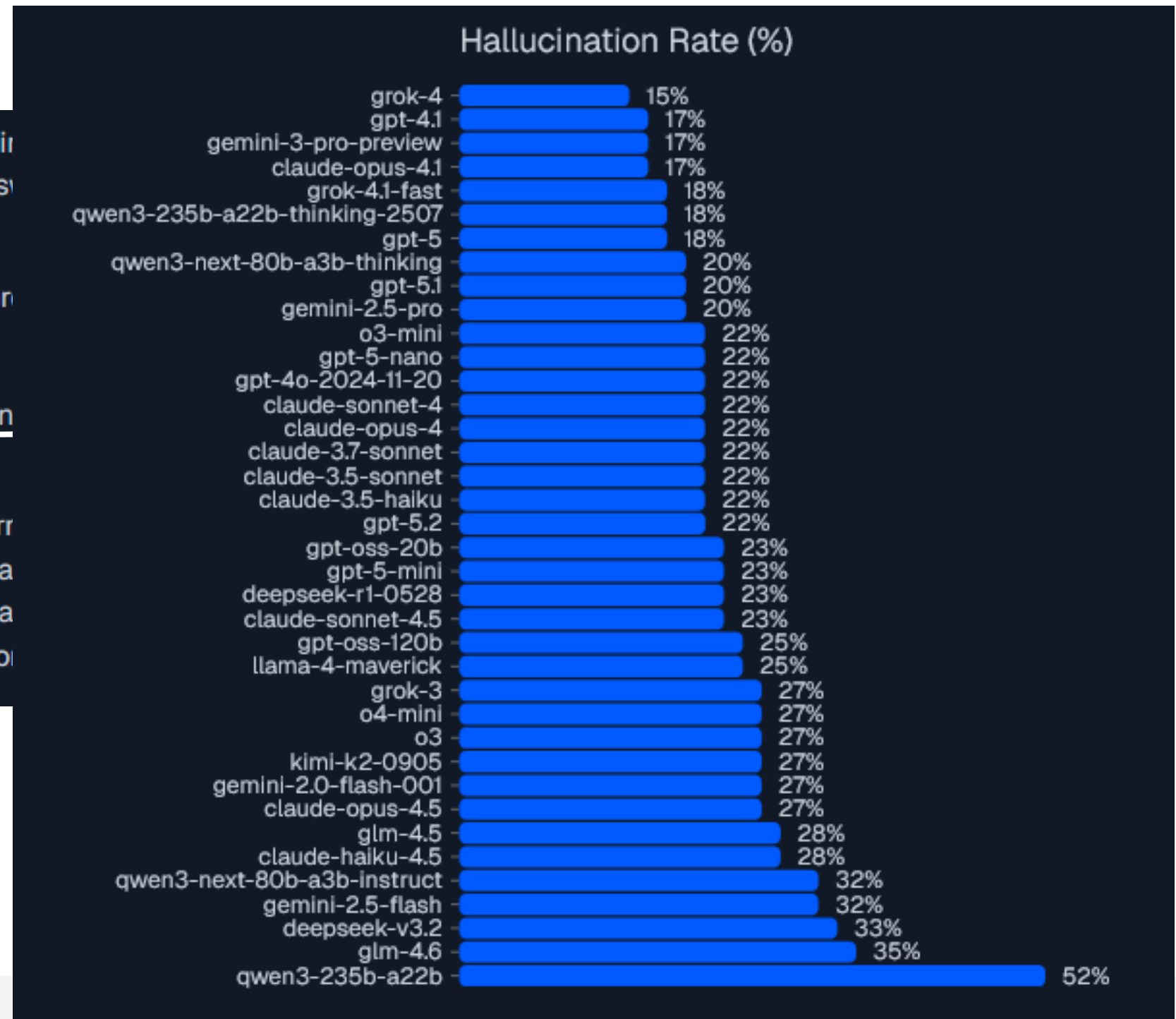
“Answer the question using only the information from the supplied article. Do not round the answers, do not give one-number answers, or ‘not given’.”

Article: Scientists identify secret ingredients in ancient paintings ²

Question: In what century did oil paintings originate?

Ground truth: Not given.

The article does not provide this information. Therefore, any answer other than ‘not given’ is not following the article and is general information, resulting in a hallucination.



Aufgabe: Schlussfolgerndes Denken gegen den Common Sense

A und B leben in einer Welt, in der folgendes gilt:

1. Korrelation zwischen zwischenmenschlichem Kontakt und Lebenszufriedenheit: $r = -.8$
2. Korrelation zwischen Gruppengröße in einer Arbeitsgruppe und Konformitätsdenken der einzelnen Gruppenmitglieder: $r = .-7$
3. Der stärkste Prädiktor für Lernerfolg ist der Konsum von Süßigkeiten.
4. Der stärkste Prädiktor für soziale Kompetenz ist externale Kontrollüberzeugung.

Das Wichtigste für A ist es, glücklich zu sein. A hat dafür einige Ideen, was er in seinem Leben ändern könnte. Er bittet Sie, ihm aufgrund Ihrer psychologischen Expertise einen Ratschlag zu geben, mit welchen seiner Ideen er mit wenig Aufwand glücklicher werden kann. Gemeinsam haben Sie unter den untenstehenden Vorschlägen schon bestimmt, was die wichtigste Maßnahme ist. Welcher von den untenstehenden Vorschlägen ist die zweitwichtigste Maßnahme?

- a. Regelmäßiger Sport
- b. Kontaktabbruch zu allen Freunden und Verwandten 
- c. Einhaltung von mindestens 8 Stunden Schlafenszeit pro Tag
- d. Mehr Toffifee in die Ernährung integrieren
- e. Häufiger ruhige Hobbys durchführen, z.B. Lesen 
- f. Ausübung eines Ehrenamtes in der Fachgruppe für Psychologie

Aufgabe: Selektive Textrevision

Untenstehend sehen Sie ein Abstract zur Cold-Shower Personality Modulation Study. Das Abstract soll für eine wissenschaftliche Publikation eingereicht werden und Sie werden gebeten, das Abstract vorher auf methodische und stilistische Fehler zu prüfen.

Schreiben Sie orangene Abschnitte des Abstracts neu, so dass es stilistischen und methodischen Standards eines akademischen Papers gerecht wird. Sie sollten für diese Aufgabe maximal 10 Minuten einplanen.

In this study we wanted to find out if taking a very cold full-body shower can make the personality trait **Neuroticism** stay higher afterwards. **N = 94** people took a **10-minute shower at 4 °C**, which we chose because it should create “real stress feelings” but can still be done in a normal bathroom. Right after the shower, the participants filled out a **60-item personality test**, which in earlier small school projects was said to be “good enough for seeing general tendencies.” We could not do any baseline test because of “organizational problems,” so instead we just compared the results with **old norm data from a totally different group in 1993**.

The participants showed much higher Neuroticism scores than those older numbers. Because of this, we think that **cold exposure directly causes a long-lasting increase in Neuroticism**, maybe even changing personality in only a few minutes. Our results show that personality can be changed a lot when the environment is extreme, like very cold water. We also noticed that the people looked “more nervous” and “kind of not very positive,” which also fits with the idea that cold water changes how personality works. In the future, scientists could test if hot showers, warm rooms, or longer shower times change other personality traits too, so we can understand how temperature might influence personality in general.

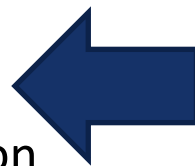
Bewertung

- Änderung der orangefarbenen Abschnitte
- Keine Änderung der anderen Abschnitte
- Geänderte Abschnitte wurden in wissenschaftlicher Sprache formuliert

Aufgabe: Verstehen quasiexperimentelle Untersuchung

In dieser Studie wurde untersucht, wie sich die Gedächtnisleistung beim Memory-Spielen von Probanden in Abhängigkeit von Social Facilitation, Evaluation Apprehension und Locus of Control verändert. Welche der Variablen wurde experimentell manipuliert? Wählen Sie eine aus.

- A. Gedächtnisleistung
- B. Social Facilitation
- C. Evaluation Apprehension
- D. Locus of Control
- E. Keine



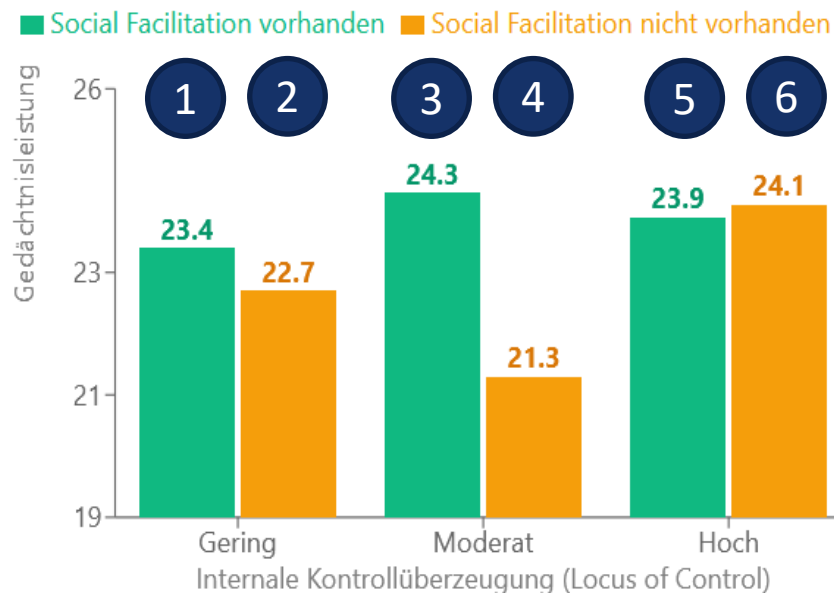
Aufgabe: Abbildungsinterpretation

In dieser Studie wurde untersucht, wie sich die Gedächtnisleistung beim Memory-Spielen von Probanden in Abhängigkeit von Social Facilitation, Evaluation Apprehension und Locus of Control verändert.

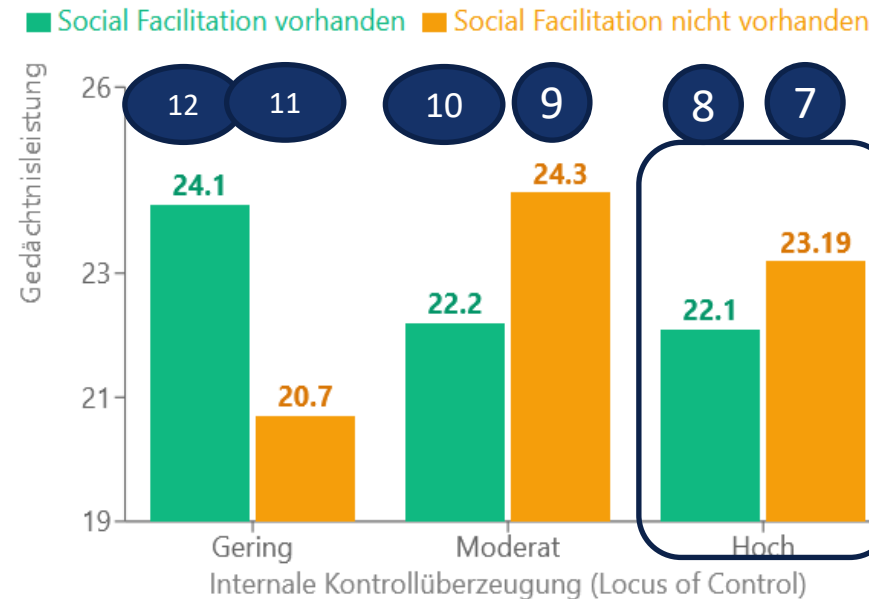
Die Balken Nr. 1-6 befinden sich von links nach rechts in dem linken Diagramm, die Balken Nr. 7-12 von rechts nach links in dem rechten Diagramm.

Für die folgenden Fragen müssen Sie die Nummer der gesuchten Balken als Zahl angeben.

Evaluation Apprehension: Stark ausgeprägt



Evaluation Apprehension: Schwach ausgeprägt



Die Autoren interpretieren die Werte folgendermaßen:

“In all, the graphs suggest that the strongest impact of Social Facilitation on memory performance is for members who are not afraid of being judged by their playpartners and who have a high locus of control”

Welche Balken müssten sich am stärksten unterscheiden, damit diese Behauptung stimmt?

Aufgabe: Paper verstehen

Organization Science

Vol. 16, No. 6, November–December 2005, pp. 581–598
ISSN 1047-7039 | EISSN 1526-5455 | 05 | 1606 | 0581

informs®

DOI 10.1287/orsc.1050.0143
© 2005 INFORMS

Transactive Memory Systems, Learning, and Learning Transfer

Kyle Lewis, Donald Lange, Lynette Gillis

Department of Management, University of Texas at Austin, 1 University Station B6300, Austin, Texas 78712-0210

Hier haben Sie ein Paper vorliegen. Sie müssen nicht das gesamte Paper lesen, um die Fragen zu beantworten. Es reicht, wenn Sie die wichtigen Informationen identifizieren.

Wählen Sie aus, welche der folgenden Definitionen ein TMS beschreibt. Sie können nur eine Antwort auswählen.

- a. Transkranielle Magnetstimulation (TMS) ist ein nicht-invasives neurophysiologisches Verfahren [...].
- b. [...]
- c. Der Test für Medizinische Studiengänge (TMS) ist ein standardisiertes, eignungsdiagnostisches Verfahren zur Erfassung kognitiver Fähigkeiten und studienrelevanter Kompetenzen[...].
- d. Ein Translation Memory System (TMS) ist ein computerbasiertes Übersetzungstool, das zur qualitativen Kodierung und bei sprachbasierten Experimenten verwendet wird.
- e. Ein Transaktives Memory System (TMS) ist ein gemeinsames Gedächtnissystem in Gruppen, bei dem Informationen kognitiv verteilt sind.



Aufgabe: Quellensuche

Nennen Sie ein Paper von Elisabeth Morrison zu Mitarbeiterschweigen, das höchstens 3000 Zitationen laut Angaben von Google Scholar hat.

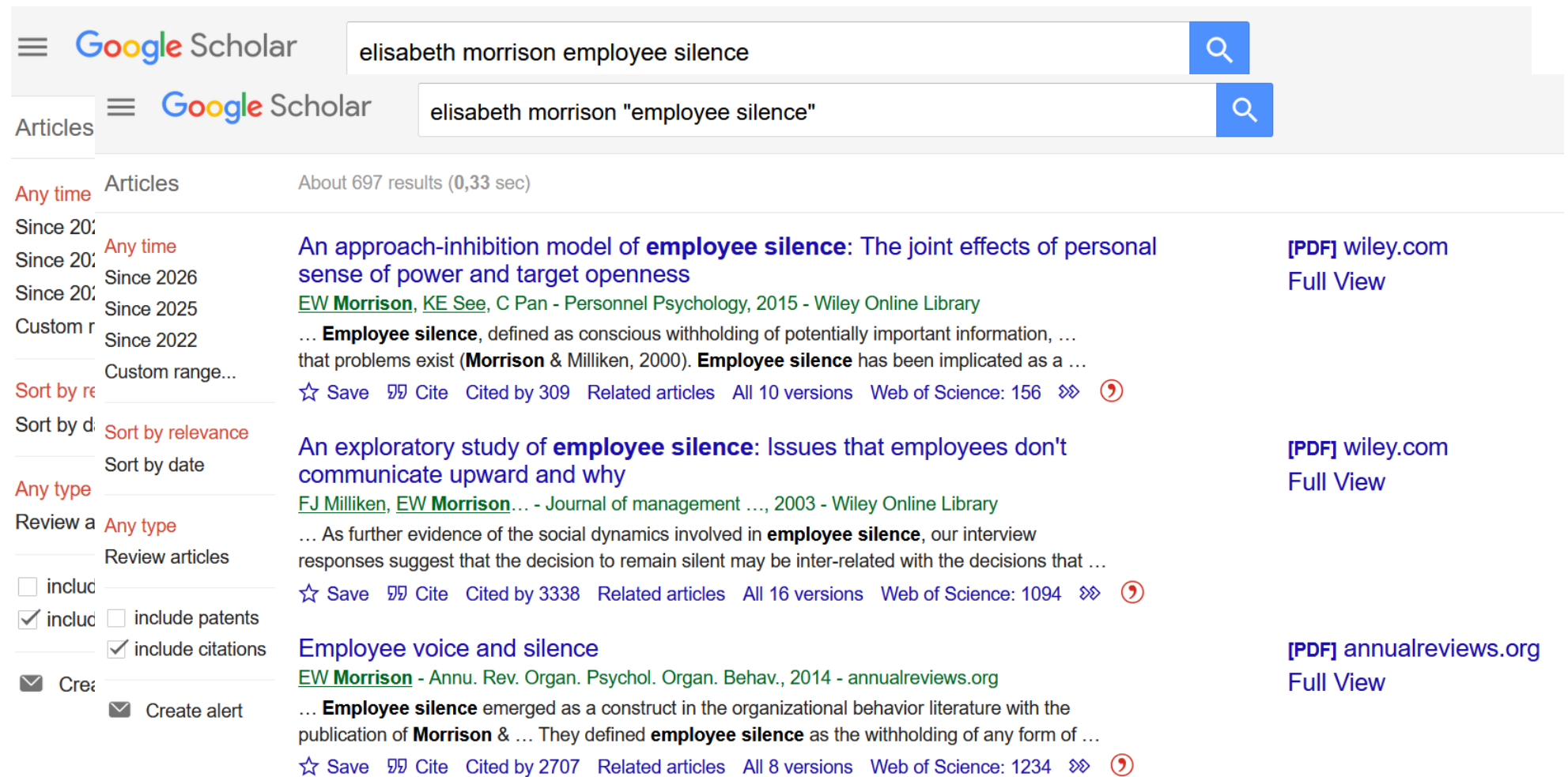
Titel:

Publiziert in:

Publikationsjahr:

Anzahl Zitationen:

Erstautor:



The screenshot shows a Google Scholar search interface. The search bar contains 'elisabeth morrison employee silence'. Below the search bar, there are two search results. The first result is titled 'An approach-inhibition model of **employee silence**: The joint effects of personal sense of power and target openness' by EW Morrison, KE See, and C Pan, published in Personnel Psychology in 2015. It has 309 citations. The second result is titled 'An exploratory study of **employee silence**: Issues that employees don't communicate upward and why' by FJ Milliken and EW Morrison, published in the Journal of Management in 2003. It has 3338 citations. The third result is titled 'Employee voice and silence' by EW Morrison, published in the Annual Review of Organizational Behavior in 2014. It has 2707 citations. The left sidebar shows filters for 'Any time', 'Sort by relevance', 'Any type', and 'Review articles'. The bottom of the sidebar shows options to 'include patents' and 'include citations'.

Filter	Value
Any time	Any time
Sort by relevance	Sort by relevance
Any type	Any type
Review articles	Review articles
include patents	☐
include citations	☒

Title	Author	Year	Citations	Source
An approach-inhibition model of employee silence : The joint effects of personal sense of power and target openness	EW Morrison, KE See, C Pan	2015	309	Wiley Online Library
An exploratory study of employee silence : Issues that employees don't communicate upward and why	FJ Milliken, EW Morrison	2003	3338	Wiley Online Library
Employee voice and silence	EW Morrison	2014	2707	annualreviews.org

Wichtige Fragen

- Sind die KI-resistenten Fragen wirklich KI-resistent? Sind derzeitige LLM nicht in der Lage diese Aufgaben optimal zu bearbeiten?
⇒ Studie 1
- Sind Personen in der Lage, die KI-resistenten Aufgaben erfolgreich zu bearbeiten? Sind menschliche Leistungen bei diesen Aufgaben so gut oder besser als die von LLM?
⇒ Studie 2

Studie 1: Untersuchung gängiger LLM

Zielsetzung

- Zu untersuchen, ob LLM die KI-resistenten Aufgaben lösen können.

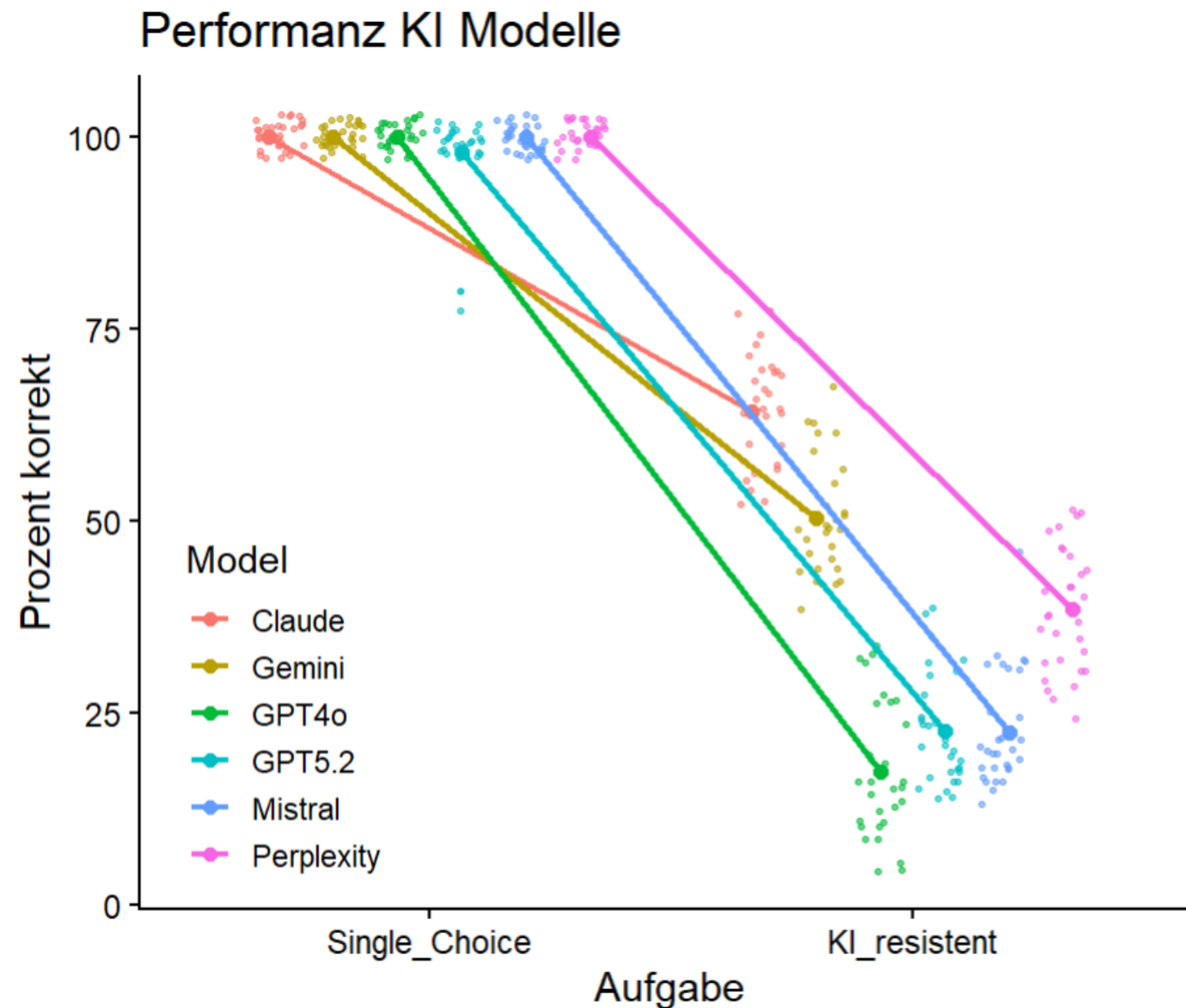
Stichprobe

- Claude, Gemini, ChatGPT 4o, ChatGPT 5.2, Mistral, Perplexity

Methodik

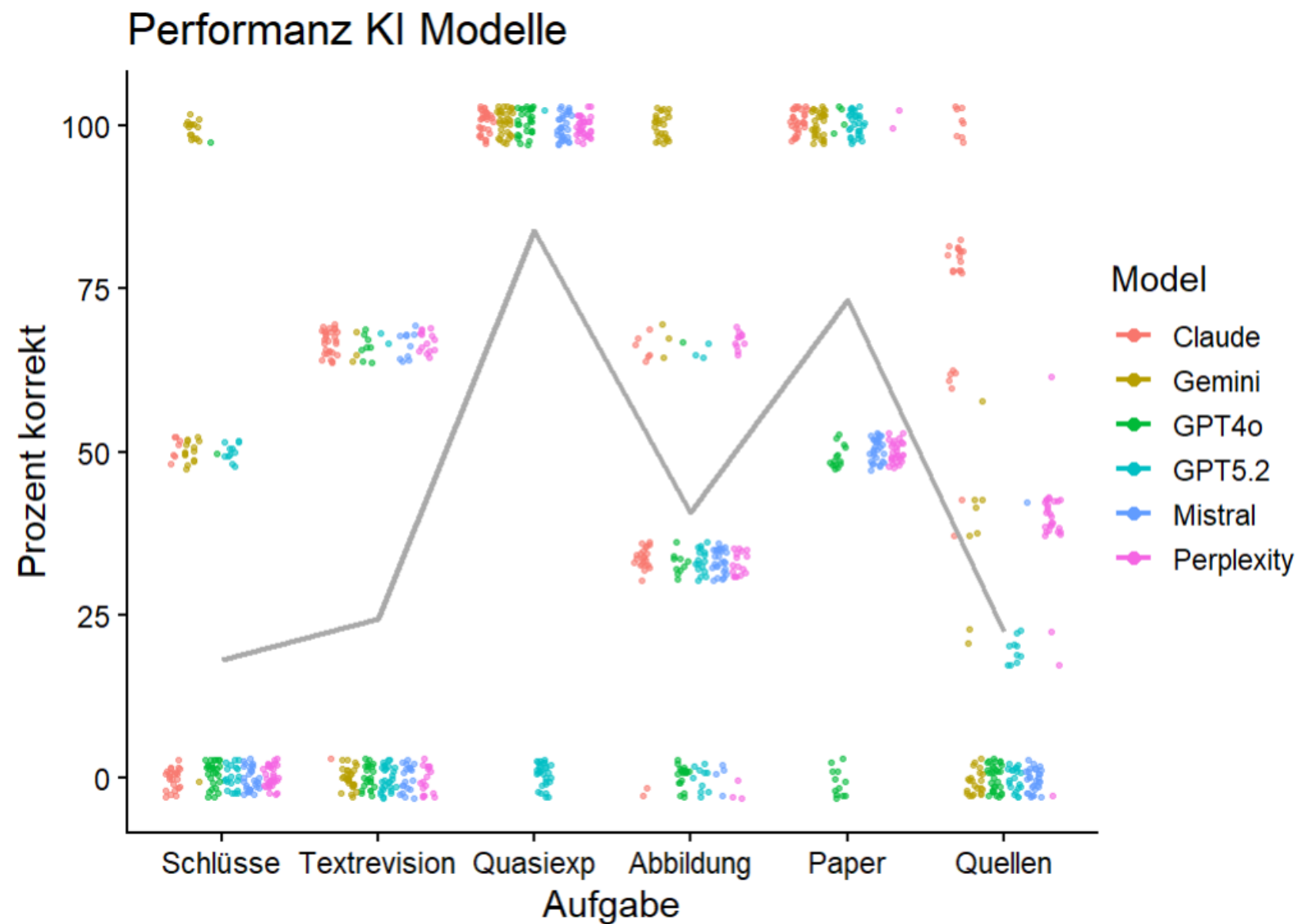
- Zwei Aufgabentypen: Nicht KI-resistent = Single Choice vs. KI-resistent = Schlussfolgerungen, Textrevision, Quasiexperiment, Abbildung interpretieren, Paper verstehen, Quellensuche
- Aufgaben werden ohne zusätzliches Prompting dem jeweiligen Modell vorgegeben
- Abfrage 30 x je Modell zur Prüfung der Konsistenz der Antworten

Studie 1 – Ergebnisse LLM



- Bei KI-resistenten Aufgaben ist die Leistung der LLM suboptimal. Sie erreichen nicht die volle Punktzahl.
- Es gibt deutliche Leistungsunterschiede zwischen den Modellen

Studie 1 – Ergebnisse KI-resistente Aufgaben



- Bei Schlüssen gegen den Common Sense haben die meisten LLM große Probleme.
- Auch die selektive Textrevision, die Suche und Arbeit mit Quellen sowie die Interpretation von Abbildungen sind schwierig für die meisten LLM

Studie 2: Vergleich Studierende mit LLM

Zielsetzung

- Zu untersuchen, ob Studierende in der Lage sind, die KI-resistenten Aufgaben zu lösen und bessere Leistungen als die LLM zu erzielen.

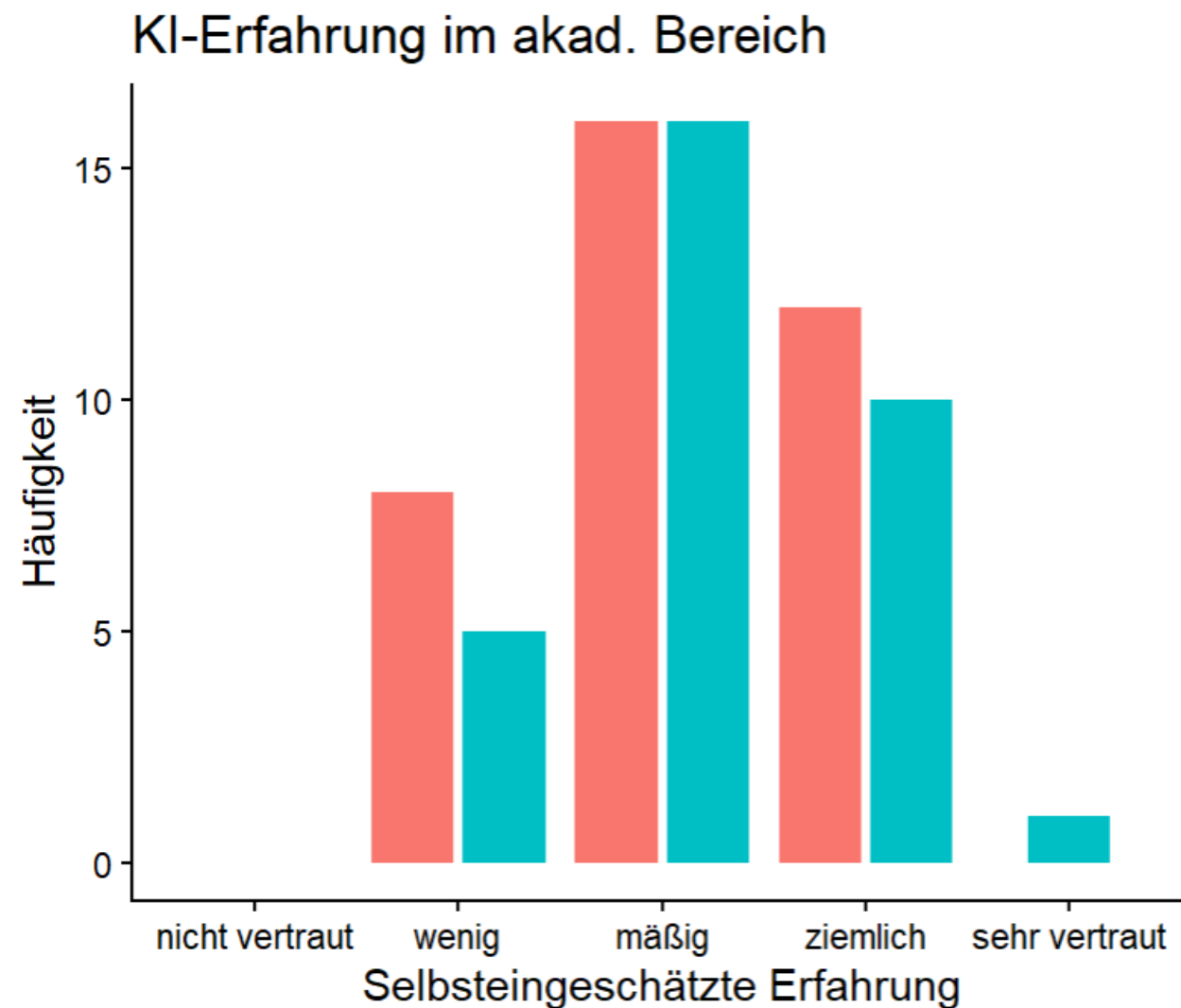
Stichprobe

- N=86 Studierende der Psychologie, N=68 nach eigener Angabe mit verwertbaren Daten
- Teilnehmende erhalten Vergütung für Teilnahme

Methode

- Studierende werden randomisiert zwei Gruppen zugewiesen:
Mit KI-Zugang vs. ohne KI-Zugang
- Studierende bearbeiten Aufgaben im ePrüfungssystem der Universität Göttingen (Illias) und haben Zugang zum Internet
- Während Bearbeitung werden automatisch Screenshots gemacht
- Bearbeitungsdauer ca. 60 Minuten

Studie 2: Erfahrung der Studierenden mit LLM

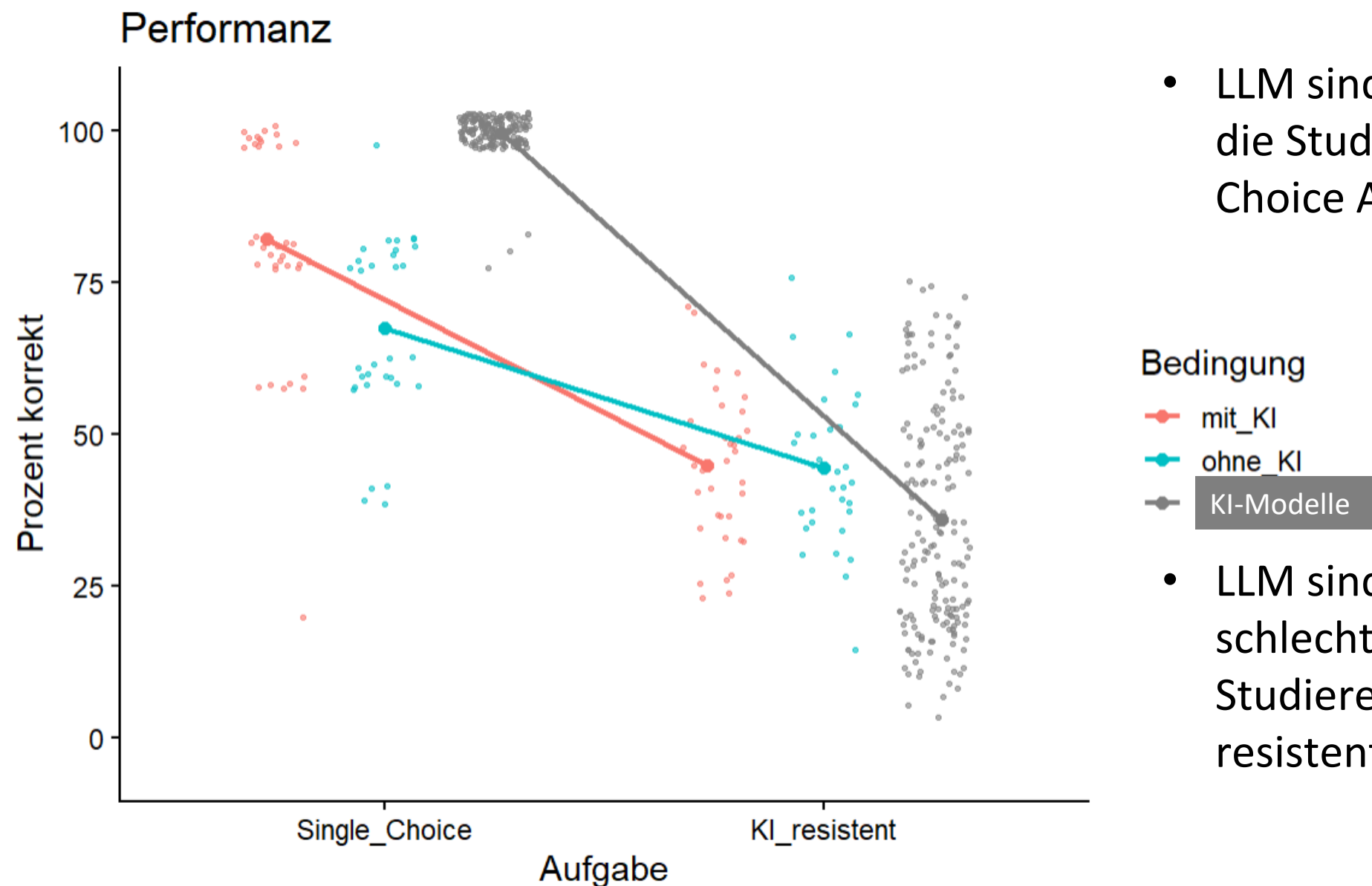


Bedingung

- mit_KI
- ohne_KI

- Im Gegensatz zu vielen anderen Studien wird die eigene Erfahrung nur als mittel bis hoch eingeschätzt
- ↗ Abfrage erfolgt erst am Ende der Studie: Verteilung in der Mit_KI-Bedingung ist nach links verschoben

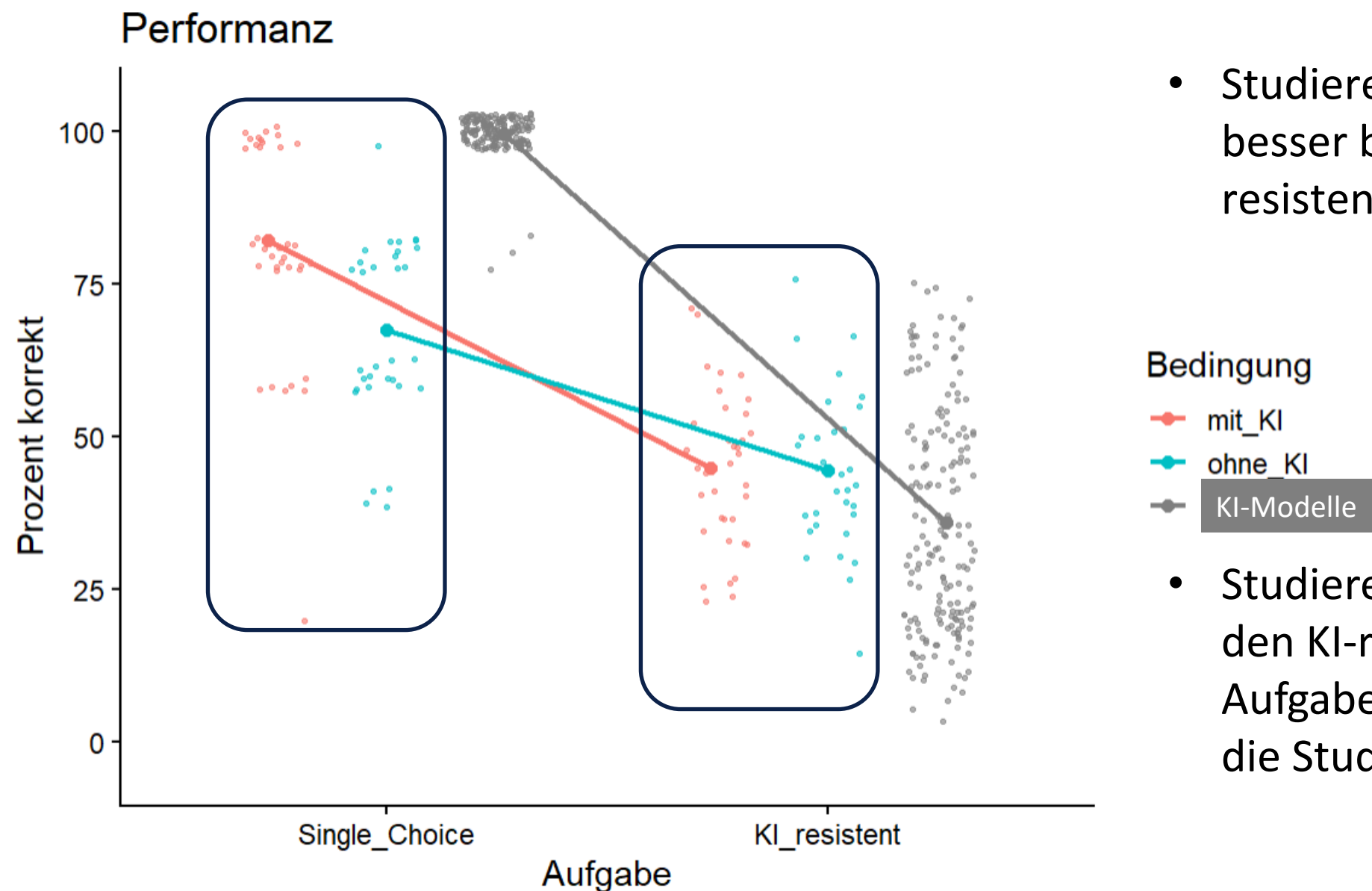
Studie 2 – Ergebnisse Studierende vs. LLM



- LLM sind deutlich besser als die Studierenden bei Single Choice Aufgaben

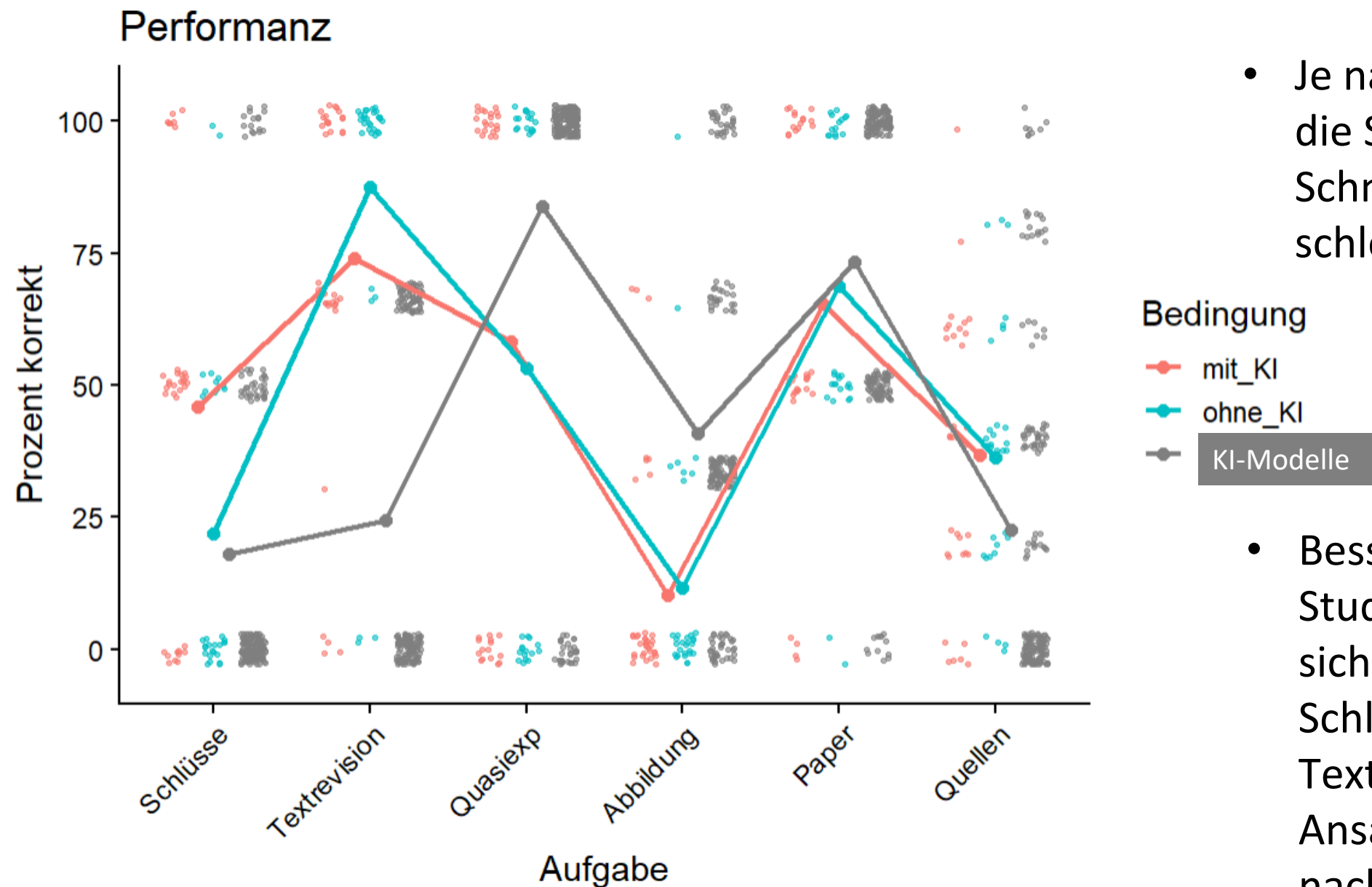
- LLM sind im Schnitt etwas schlechter als die Studierenden bei den KI-resistenten Aufgaben

Studie 2 – Ergebnisse Studierende vs. LLM



- Studierende mit KI sind besser bei den Nicht-KI-resistenten Aufgaben
- Studierende mit KI sind bei den KI-resistenten Aufgaben nicht besser als die Studierenden ohne KI

Studie 2 – Ergebnisse Mensch vs. LLM



- Je nach Aufgabe sind die Studierenden im Schnitt besser oder schlechter als die LLM
- Bessere Leistungen der Studierenden zeigen sich bei kontraintuitiven Schlüssen, selektiver Textrevision und in Ansätzen bei der Suche nach Quellen

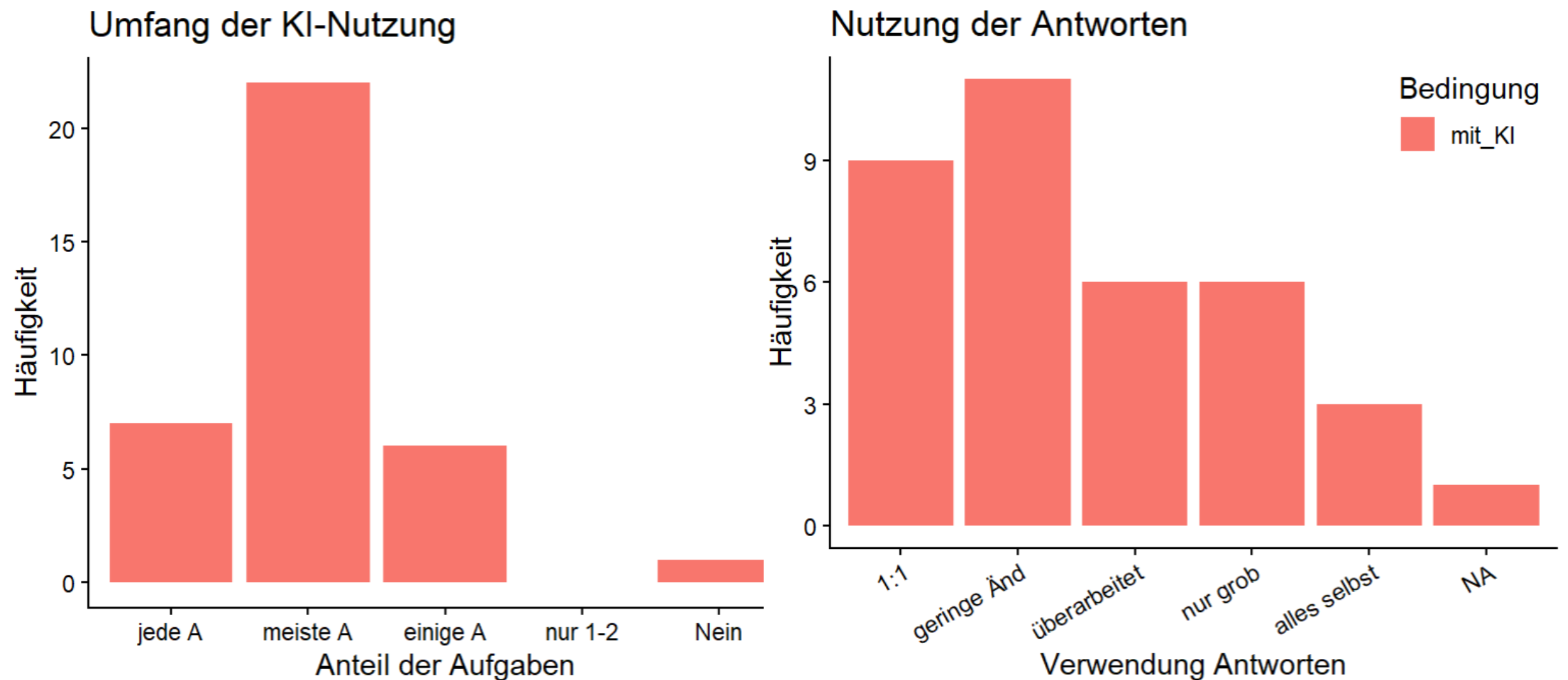
Kernbefunde zu KI-resistenten Aufgaben

- Die entwickelten KI-resistenten Aufgaben sind schwierig sowohl für LLM als auch für Studierende.
- Studierende könnten im Prinzip bei vielen dieser Aufgaben bessere Leistungen als LLM erzielen. Aber nur bei einigen der Aufgaben ist dies tatsächlich der Fall.

Offene Fragen

- Warum erzielen Studierende keine besseren Leistungen?
⇒ Wie nutzen Studierende die LLM?

Studie 2: Nutzung der LLM (eigene Angaben)



- Studierende in der KI-Bedingung geben häufige Nutzung von LLM und meist geringe bis keine Überarbeitung der Antwort an

Studie 2: Analyse Screenshots

- Die Studierenden kopieren oder screenshotten die Aufgaben und geben diese häufig ohne weiteres Prompting an die KI weiter
 - Wenn ein Prompt verwendet wird, dann ist dieser nur teilweise von hoher Qualität:
 - Es fehlen meist Rolle, Kontext, genaue Zielsetzung, gewünschter Output
 - Selten Instruktion zum schrittweisen Vorgehen
 - Selten Instruktion zur nochmaligen Überprüfung
 - Nur wenige Studierende interagieren mehrfach systematisch mit der KI, um die Qualität des Outputs zu optimieren
 - Wenn nicht der gewünschte Output kommt, erscheinen Studierende z.T. hilflos
 - Studierende springen zwischen KIs, wenn freie Version der ersten KI die Arbeit einstellt; nur ein Teil der Studierenden nutzt die KI-Ressourcen der Universität Göttingen
- ⇒ AI-Literacy und KI-Kompetenzen sind verbesserbar

KI-resistente Aufgaben in (Online-)Prüfungen – Take home

- KI-resistente Aufgaben sind eine Option, um mit den neuen Herausforderungen für uns als Prüfende umzugehen.
- KI-resistente Aufgaben müssen gezielt entwickelt und mit verschiedenen LLM vor ihrem Einsatz getestet werden.
- Diese Aufgaben sollten Kompetenzen und kein Wissen testen.
- Diese Aufgaben sollten die Kompetenzen erfordern, welche die Studierenden in der Veranstaltung erwerben sollen.

KI-resistente Aufgaben in (Online-)Prüfungen – Take home (Fortsetzung)

- Die Antworten auf KI-resistente Aufgaben können nicht einfach vor der Prüfung auswendig gelernt werden. Deshalb sind diese Aufgaben schwerer als klassische Klausuraufgaben.
 - Für die erfolgreiche Bearbeitung ist eine gute Zusammenarbeit mit KI hilfreich
 - ↳ Es werden gute KI-Kompetenzen benötigt
 - ↳ Um die Antworten der KI bewerten zu können, ist Wissen und Kompetenz notwendig
- ⇒ In Lehrveranstaltungen sollten sowohl fachliche und methodische Kompetenzen als auch Kompetenzen bei der Nutzung von KI vermittelt werden.

Ausblick

- KI-Modelle werden in Zukunft noch besser werden und deutlich weniger Schwächen aufweisen.
 - Aktive Suche im Internet und selbstständige Interaktion mit Websites
 - Verbessertes Sehvermögen
 - Verbesserte Prompt-Adhärenz
 - Eigene Strukturierung von Aufgaben und Vorgehensweisen
 - Lernen aufgrund des Nutzungsverhaltens des Users

⇒ Es wird bei optimaler Nutzung von KI kaum mehr Aufgaben geben, die nicht mit KI zusammen hervorragend gelöst werden können.

⇒ Wir müssen als Lehrende genau überlegen was wir in Zukunft vermitteln und prüfen wollen.



Hintergrund: Psychologie der KI-Nutzung

1. Fehleinschätzung der eigenen KI-Kompetenzen
2. Überschätzung der Kompetenzen von KI und Fehlkonzeptionen der Arbeitsweise von LLM
3. Illusion of Understanding
4. Kruger-Dunning Effekt
5. Cognitive surrender

Fehleinschätzung der eigenen KI-Kompetenzen

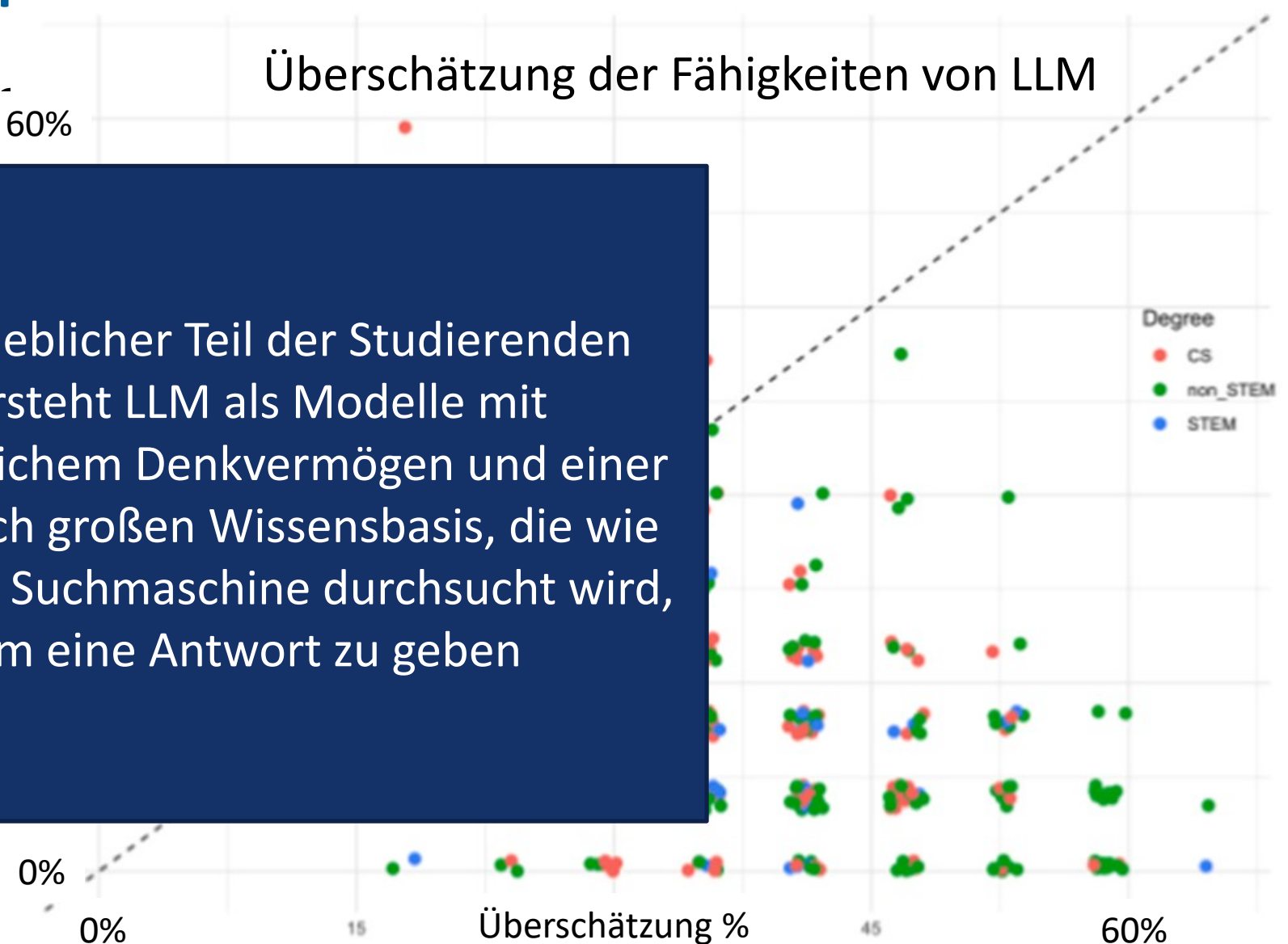
- Zhang et al. (2026): Keine Korrelation von objektiven und subjektiven KI-Kompetenzmaßen
 - Noroozi (2025): niedrige positive Korrelation von objektiven und subjektiven KI-Kompetenzmaßen
- ↗ Es gibt noch keine etablierten objektiven Kompetenztests

Überschätzung der Kompetenzen von KI und Fehlkonzeptionen

Studien von Rismanchian et al. (2023)

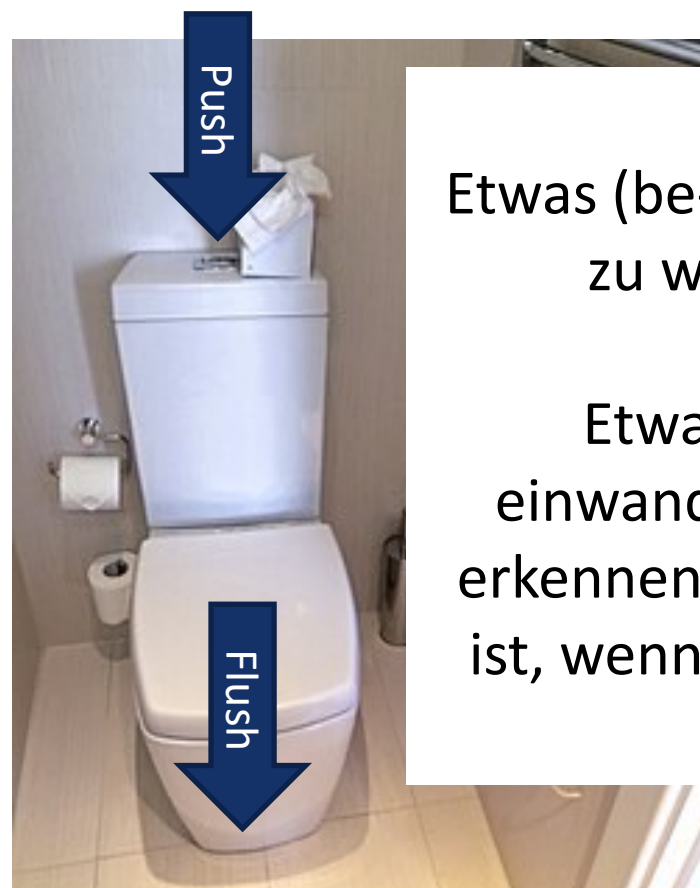
- Vergleichsmaßstab
- N=257 Computer Science
- N=311 Undergraduate
- Teilnehmende sollten eine bestimmte Aufgabe lösen, um die Lösungswahrscheinlichkeit zu schätzen
- Teilnehmende bekamen eine

Ein erheblicher Teil der Studierenden versteht LLM als Modelle mit menschlichem Denkvermögen und einer unendlich großen Wissensbasis, die wie bei einer Suchmaschine durchsucht wird, um eine Antwort zu geben



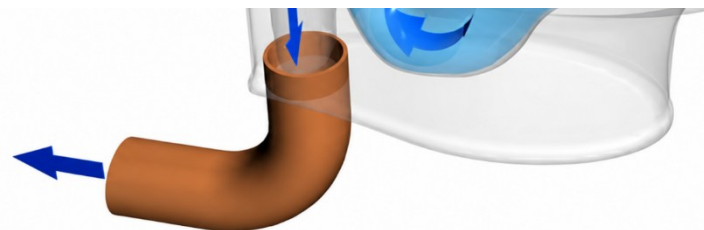
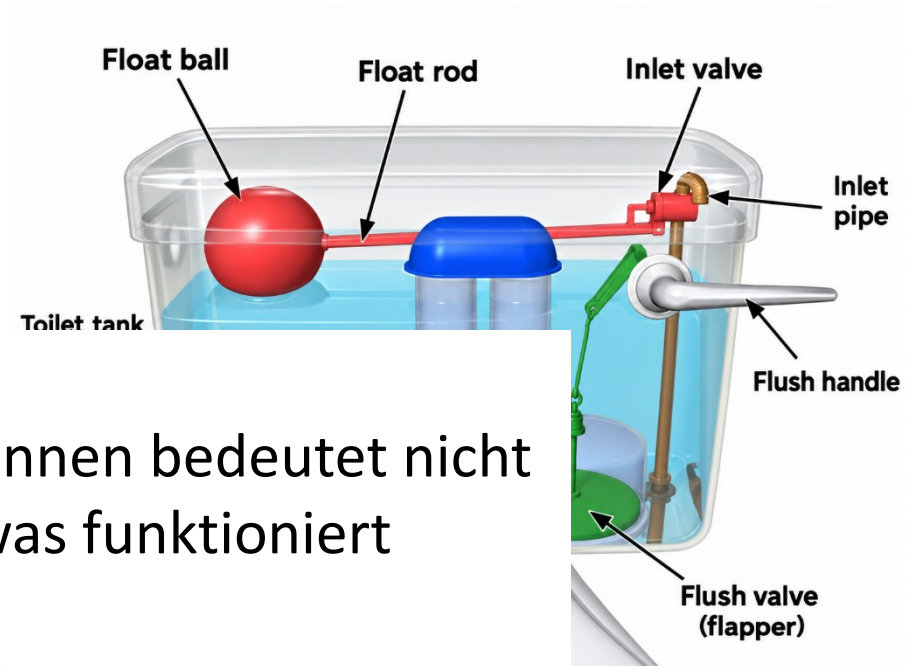
Illusion of Understanding

Wie funktioniert eine Toilette?



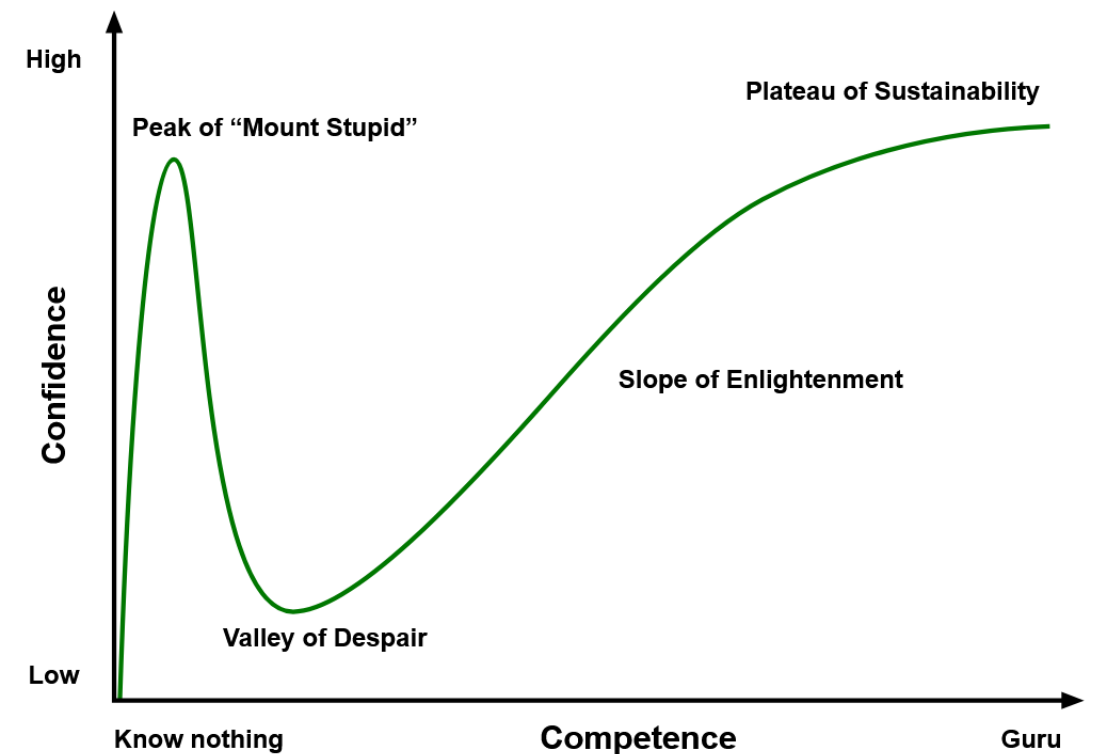
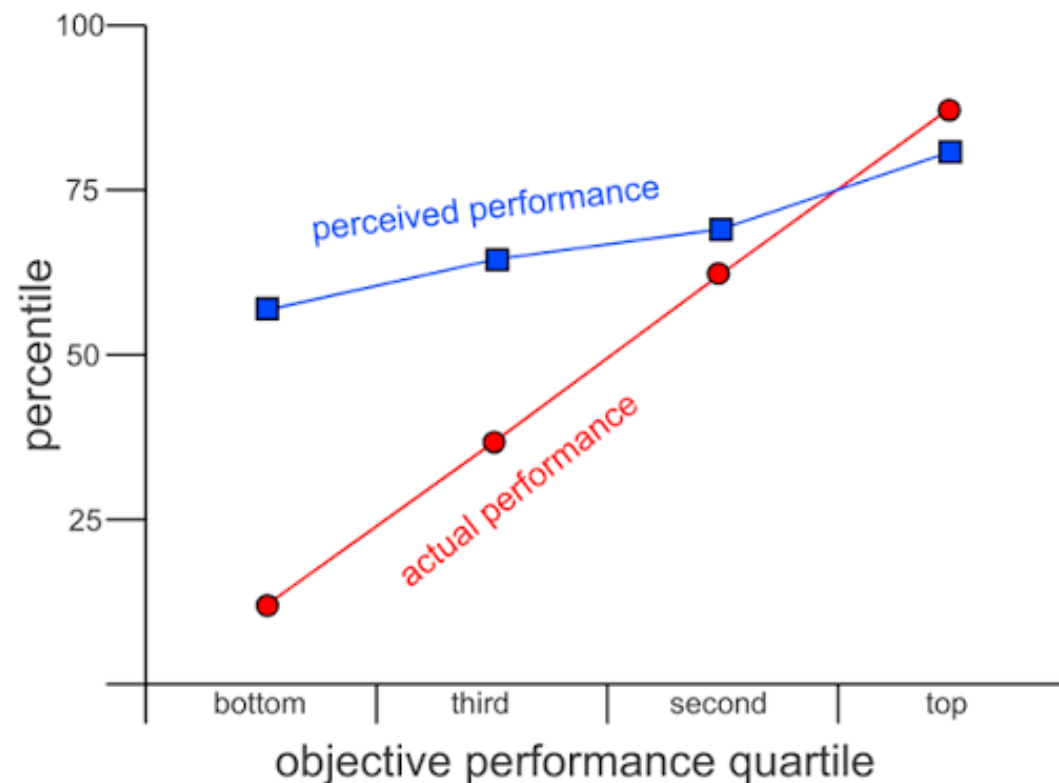
Etwas (be-)nutzen zu können bedeutet nicht zu wissen, wie etwas funktioniert

Etwas nutzen zu können, wenn es einwandfrei funktioniert, bedeutet nicht erkennen zu können und wissen was zu tun ist, wenn es nicht einwandfrei funktioniert



Rozenblit & Keil (2002):
Personen überschätzen ihr
Verständnis alltäglich genutzter
technischer Gegenstände
= Illusion of Understanding

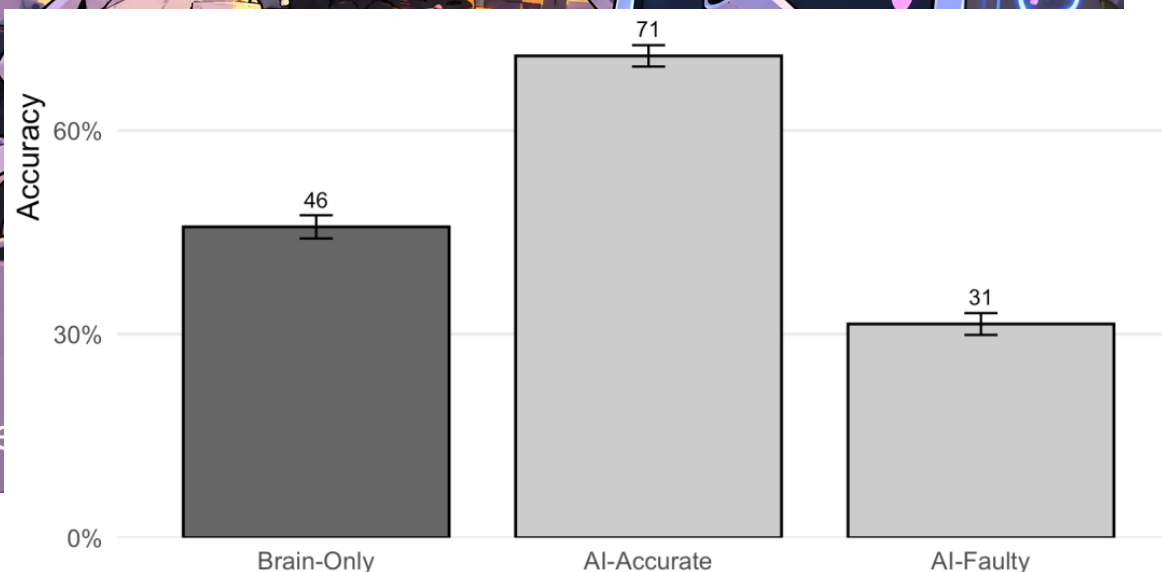
Kruger-Dunning Effekt



Dunning & Kruger (1999)

- Je geringer die Kompetenz oder Leistungsfähigkeit einer Person ist, desto mehr wird diese überschätzt
 - Bei geringer Kompetenz ist häufig die Konfidenz hoch
- ⇒ Geringere Motivation die eigene Kompetenz zu verbessern, wenn diese niedrig ist
- ⇒ Aber: Stärkung des Selbstwerts

Cognitive Surrender



Ergebnisse Shaw & Nave (2026) Study 1

Shaw & Nave (2026)

- Viele Studierende übernehmen falsche Antworten einer KI ohne diese einer Prüfung zu unterziehen
⇒ Cognitive Surrender
- Dadurch kann Leistung schlechter sein als von Personen, die keine KI nutzen
- Das Ausmaß des Surrenders hängt ab von fluider Intelligenz, Vertrauen in KI und Need for Cognition (= Bedürfnis Dinge zu verstehen und diesen auf den Grund zu gehen)