LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CHAIR OF PHYSICS EDUCATION – FACULTY OF PHYSICS
Didaktik der Physik

Evidence of Impact and Interpretational Limits of Generative AI in STEM education

– A systematic Review and Meta-Analysis

Stefan Küchemann



AI in STEM Education - Linz

Acknowledgment

- **AI in Physics Education, LMU:**
  - Chiara Boolzen,
  - Nina Peltzer
  - Prof. Dr. Jochen Kuhn
  - Dr. Karina Avila,
  - Steffen Steinert,
  - Yavuz Dinc,
  - Natalia Revenga,
  - Dr. Verena Ruf,
  - Dr. Niklas Stausberg,
  - Caroline Wermann,
- **TU Dresden**
  Jun.-Prof. Dr. Salome Felgr

Federal Ministry of Research, Technology and Space

European Commission

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Content

1. Introduction – Example of use of generative AI in STEM education
2. Methods – Research Questions and search within the PRISMA Framework
3. Results – Overall effect, Moderator analysis, Meta-regression & Coincidence Analysis
4. Hypothetical Framework
5. Implications for STEM teaching practice

Boolzen, C., Peltzer, N., Flegr, S., Kuhn, J., Küchemann, S., (2026), Evidence of Impact and Interpretational Limits of Generative AI in STEM education - A Systematic Review and Meta-Analysis on Cognitive Learning Outcomes, arXiv:submit/7369304

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

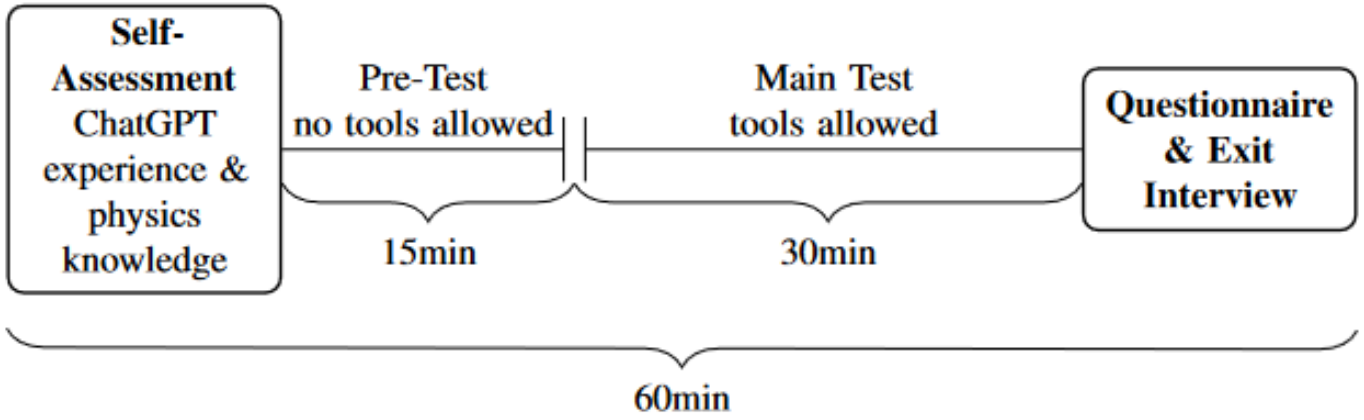
Didaktik der Physik

Introduction

Two example uses of generative AI in STEM education

Possible support of physics students during problem solving

N = 27 Physics- and Engineering students



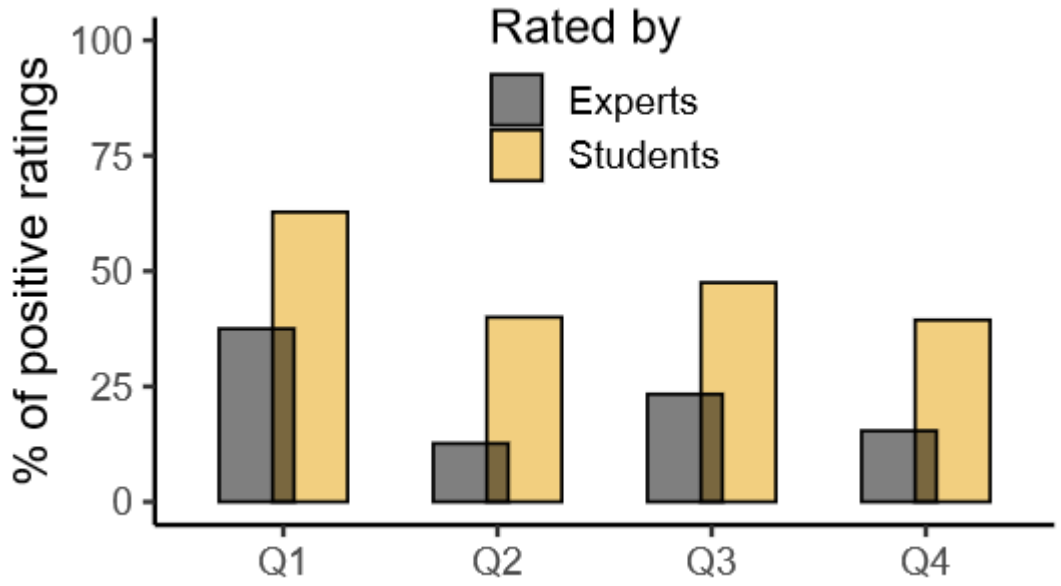
HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

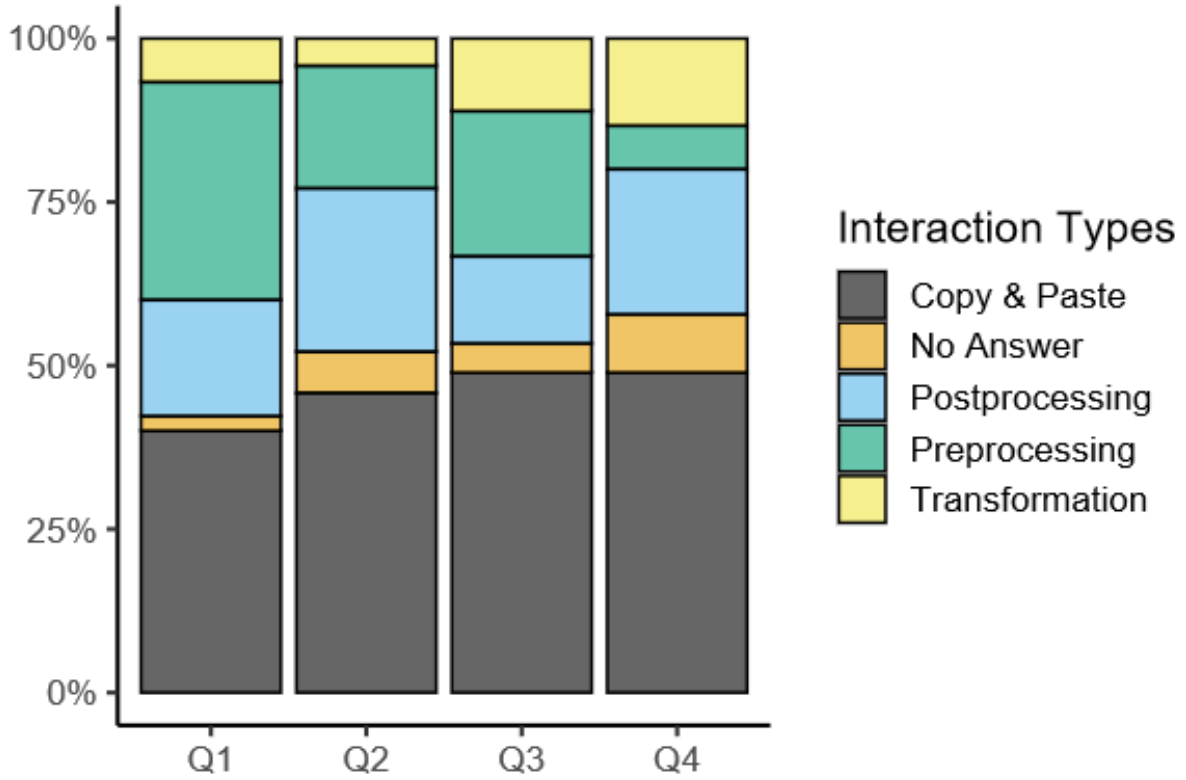
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA240195

Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education

Lars KRUPP [a,b,1], Steffen STEINERT [c] Maximilian KIEFER-EMMANOUILIDIS [a,b] Karina E. AVILA [b] Paul LUKOWICZ [a,b] Jochen KUHN [c] Stefan KÜCHEMANN [c] and Jakob KAROLUS [a,b]

Possible support of physics students during problem solving





Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In HHAI 2024: Hybrid Human AI Systems for the Social Good (pp. 199-212). IOS Press.

Cognitive activity during essay writing with ChatGPT

Comparison of three groups:

1. ChatGPT group
2. Google group
3. No support

Results:

1. Systematical differences in neural connectivity patterns between groups
2. Brain connectivity scaled down with increasing level of external support: No support group reached highest connectivity, followed by Google group and the lowest connectivity group was reached by ChatGPT group
3. the ChatGPT group also reached lowest scores in ownership and content retention

Learning requires the active construction of cognitive schemata. This work suggests that students working with ChatGPT do invest relevant effort in constructing schemata

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., ... & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. arXiv preprint arXiv:2506.08872.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Methods

Research Questions and search within the PRISMA Framework

Research questions

1) To what extent does interaction with GAI enhance learning effectiveness?
    a) What moderator variables influence the interaction between learners and GAI systems?

2) What types of challenges do learners encounter when interacting with GAI systems?

3) What interventions have been reported to promote successful interactions between learners and GAI systems?

Boolzen, C., Peltzer, N., Flegr, S., Kuhn, J., Küchemann, S., (2026), Evidence of Impact and Interpretational Limits of Generative AI in STEM education - A Systematic Review and Meta-Analysis on Cognitive Learning Outcomes, arXiv:submit/7369304

What is generative AI?

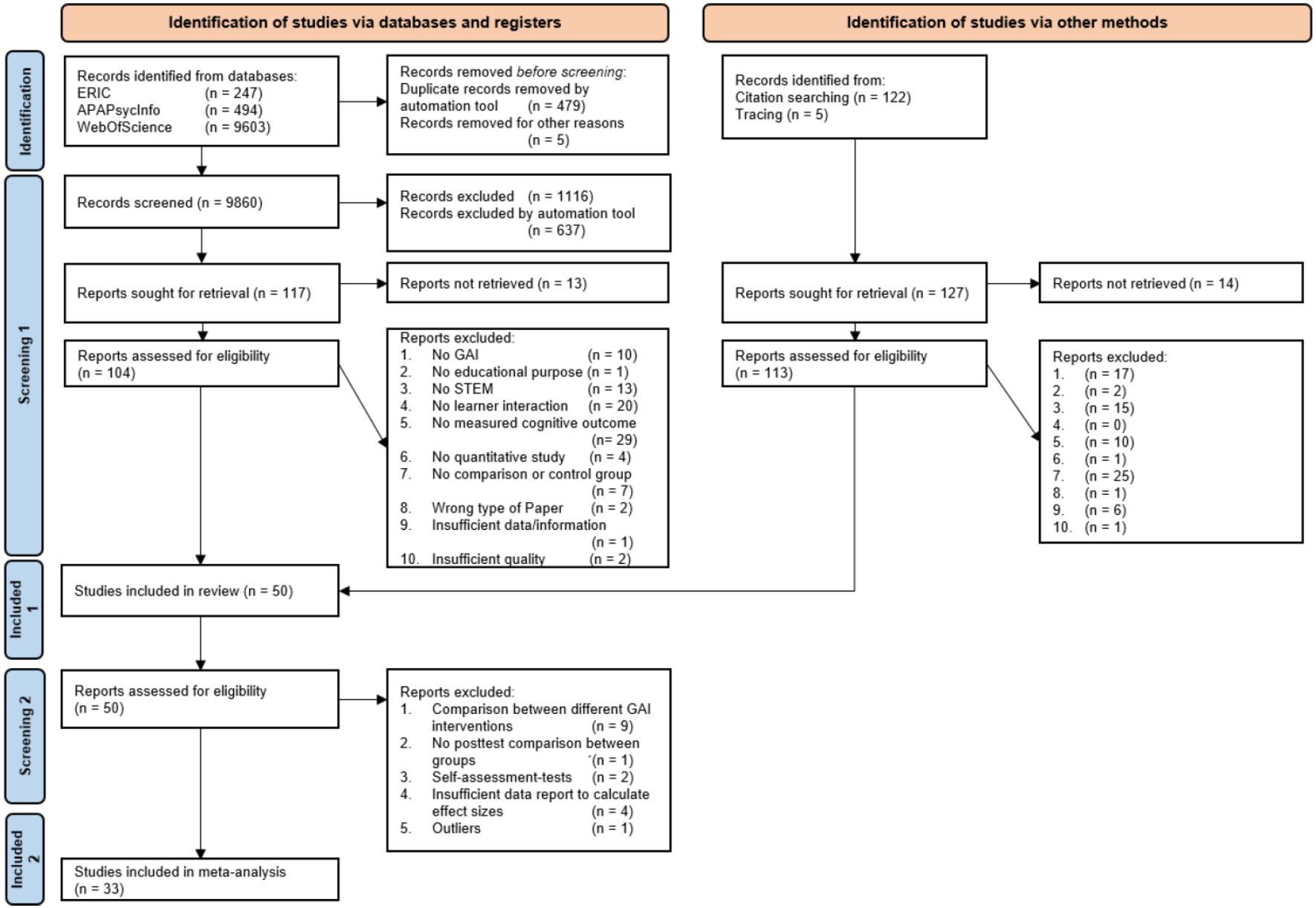
- Diverging social and technical understaning of the term
- No universally agreed definition – „GAI is a distinct construct, which contributes to definitional pluralism in technical and public discourses" Ronge et al. (2025)
- **Definition by NIST:**
  "The class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content" (NIST Computer Security Resource Center (CSRC), 2025)
- Possible Categorization of GAI systems (Boolzen & Peltzer et al., 2026):

Input \ Output	High (open-ended output)	Low (constrained output)
High (open-ended input)	HH — Generative. (e.g., ChatGPT (OpenAI, 2022), Stable Diffusion (Stability AI, 2022), GitHub Copilot (GitHub and OpenAI, 2021))	HL — Generative. (e.g., Google Cloud Natural Language–Content Classification (Google Cloud, 2016), Dialogflow (Google Cloud, 2017), Rasa NLU–intent Classification (DIET) (Rasa, 2020))
Low (constrained input)	LH — Generative. (e.g., Canva Magic Write (Canva, 2022), Duolingo (Duolingo, Inc., 2012), Khan Academy (Khan Academy, 2008)).	LL — Non-generative. (e.g., ALEKS (ALEKS Corporation, 1993), kahoot (Kahoot! ASA, 2013)).

Boolzen, C., Peltzer, N., Flegr, S., Kuhn, J., Küchemann, S., (2026), Evidence of Impact and Interpretational Limits of Generative AI in STEM education - A Systematic Review and Meta-Analysis on Cognitive Learning Outcomes, arXiv:submit/7369304
NIST Computer Security Resource Center (CSRC) (2025) Generative artificial intelligence. https://csrc.nist.gov/glossary/term/generative-artificial-intelligence, glossary entry, retrieved November 12, 2025
Ronge R, Maier M, Rathgeber B (2025) Towards a definition of generative artificial intelligence. Philosophy & Technology 38(1):31

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Studies included



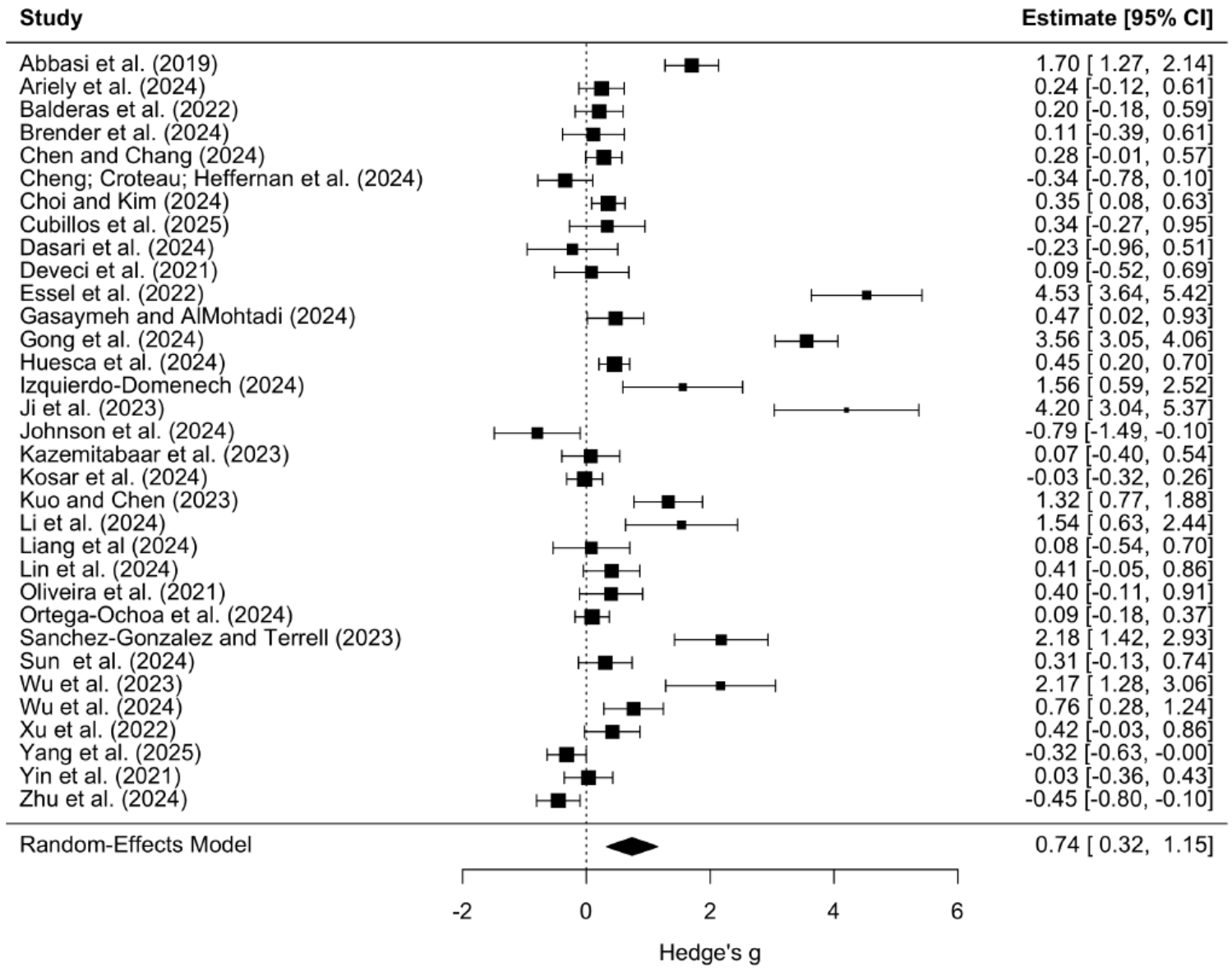
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Results

Overall effect, Moderator analysis, Meta-regression & Coincidence Analysis

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Overview of results (Boolzen and Peltzer et al., 2026)



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Moderator Analysis

Moderator	Level	k	Z	Q_B	p-value	p (BH)	g (95% CI)
domain		33		0.03	0.861	0.876	
	IT	18	2.45		0.014*		0.706 [0.140, 1.272]
	OTH	15	2.43		0.015*		0.782 [0.150, 1.41]
educational		32		1.18	0.555	0.777	
level	Uni	19	1.43		0.230		1.556 [–0.984, 4.10]
	K12	13	1.43		0.150		0.488 [–0.177, 1.15]
GAI system		33		0.03	0.853	0.876	
	GPT	15	2.19		0.029*		0.697 [0.072, 1.321]
	OTH	18	2.66		0.008**		0.777 [0.205, 1.348]
adapted		33		0.02	0.876	0.876	
system	no	13	2.03		0.042*		0.698 [0.025, 1.372]
	yes	20	2.78		0.005**		0.767 [0.227, 1.307]
learning		33		5.99	0.014*	0.065	
outcome	K	17	4.36		$1.28 \cdot 10^{-5}$***		1.213 [0.668, 1.759]
	S	16	0.88		0.375		0.248 [–0.300, 0.796]
GAI type		33		0.06	0.806	0.876	
	HH	20	2.52		0.012*		0.697 [0.155, 1.239]
	HL	13	2.36		0.018*		0.805 [0.136, 1.475]
thinking		33		0.49	0.485	0.754	
skill	HOT	11	2.55		0.011*		0.954 [0.220, 1.688]
	LOT	22	2.44		0.015*		0.636 [0.126, 1.146]
interaction		28		1.58	0.208	0.416	
duration	< 4 w	14	1.38		0.169		0.456 [–0.193, 1.11]
	> 4 w	14	3.17		0.002**		1.045 [0.398, 1.69]

Meta-Regression and Coincidence Analysis

Meta regression (Moderators: learning outcome and intervention duration):
The main effect model (learning outcome + intervention duration) explained a large portion of the variance in effect sizes (QM(3) = 9.44, p = 0.024) and reduced cross-study heterogeneity by R2 = 21.54%.

Coincidence Analysis
Identification combinations of intervention duration and learning outcome which could be a necessary condition for large effects.
The analysis yielded neither atomic nor complex solution terms for the outcome “large effect”.

Meta-Meta-Analysis – Effects of AI & LLMs

- We challenge these findings by a study-level meta-meta-analysis of 1,840 effect sizes from 67 meta-analyses.
- The results show strong evidence of severe publication bias and extreme between-study heterogeneity.
- Publication bias-adjusted analyses reveal model-averaged effects approximately one-third the magnitude of unadjusted estimates (standardized mean difference, SMD= 0.196, 95% credible interval from 0.000 to 0.323),
- Along with wide prediction intervals spanning both large negative and positive effects (-1.521 to 1.908).
- Subgroup analyses fail to identify specific moderators that yield consistent benefits.

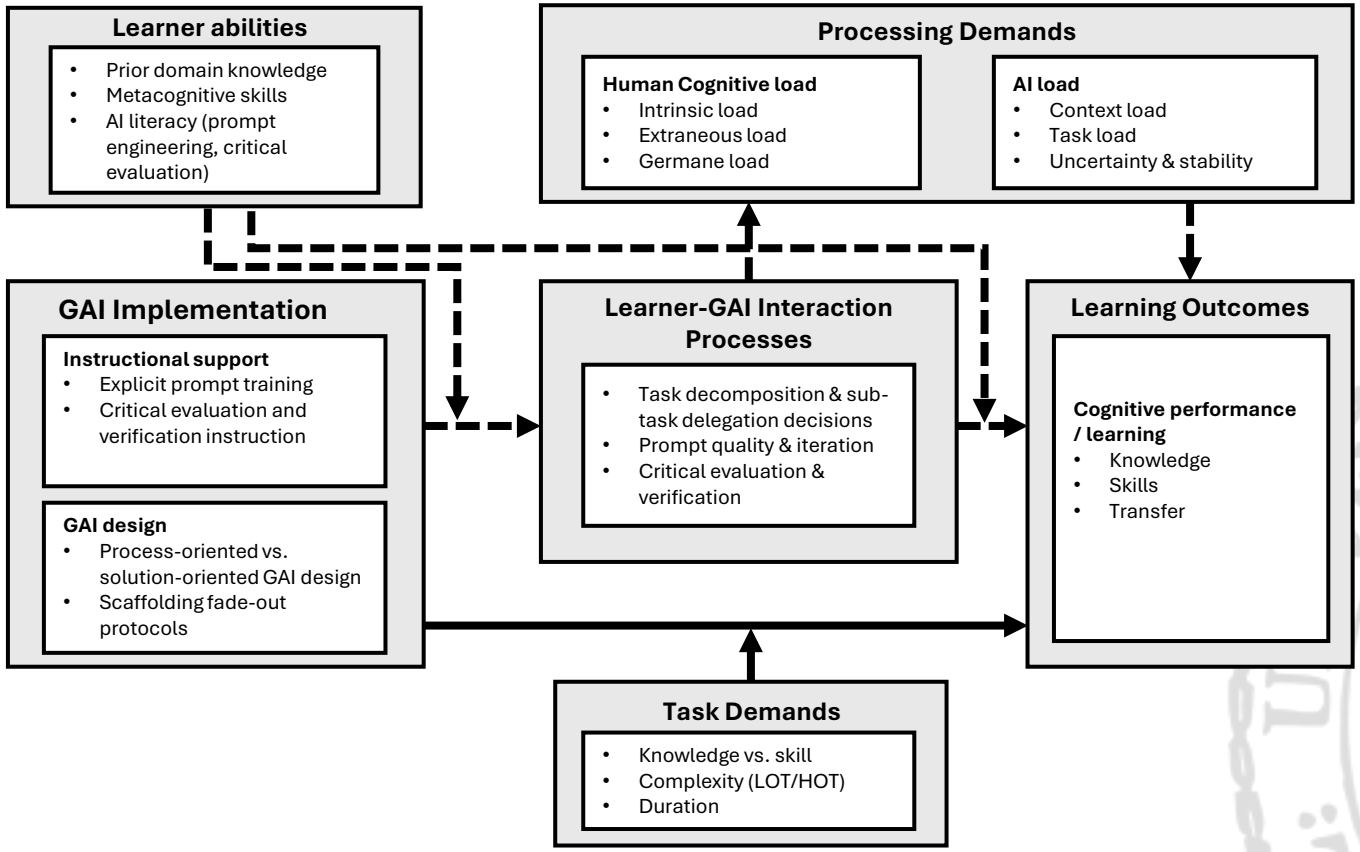
Bartoš, F., Bujak, O., Martinková, P., & Wagenmakers, E. J. Effect of Artificial Intelligence on Learning: A Meta-Meta-Analysis.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Hypothetical Framework

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Hypothetical Framework



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Implications for STEM Teaching Practice

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

What do the results mean for STEM teaching?

Key finding: Generative AI in STEM education shows medium to large effects to cognitive learning outcomes (g = 0.74). It does not matter which genAI-tool is used, but how it is implemented.

Tool selection & Configuration

- No systematic difference between ChatGPT and other LLMs
- The LLM provide is secondary but accessibility, data protection standards and a match with the learning goal are important
- Purposeful inclusion of errors of genAI as learning goals

Learning support

- Explicitely teaching of prompt engineering or scaffolding support (e.g., worked examples)
- Instruct verification strategies (e.g., dimensional analysis, plausability checks)
- Support in task delegation decisions to avoid cognitive overload

Task design & activities

- Activities that target knowledge outcomes benefit stronger from genAI than competences
- Constructive & interactive activities (ICAP & ISAR Framework)
- Iterative use of genAI: students activity first – support by genAI – revision/extension

Duration

- Longer interventions (>4 weeks) show larger effects (g = 1.05)
- Recommended sequence:
  1. Scaffolding support for an effective use of genAI
  2. Guided use of genAI
  3. Enhanced autonomy

Thank you for your attention!

AOR Dr. Stefan Küchemann
Group Leader AI in Physics Education

Ludwig-Maximilians-Universität München / *LMU Munich*
Fakultät für Physik / Lehrstuhl Didaktik der Physik
E-mail: s.kuechemann@lmu.de