Molerov, Dimitri; Federiakin, Denis; Zlatkin-Troitschanskaia, Olga; Shenavai, Kevin; Trierweiler, Lukas; Nagel, Marie-Theres

The relationship between AI-chatbots use, student assessment performance and learning outcomes in higher education

Unterrichtswissenschaft 54 (2026) 3, S. 325-356

Unterrichtswissenschaft (2026) 54:325–356
https://doi.org/10.1007/s42010-026-00242-2

THEMENTEIL

The relationship between AI-chatbots use, student assessment performance and learning outcomes in higher education

Dimitri Molerov · Denis Federiakin · Olga Zlatkin-Troitschanskaia · Kevin Shenavai · Lukas Trierweiler · Marie-Theres Nagel

Received: 26 May 2025 / Revised: 15 December 2025 / Accepted: 15 January 2026 / Published online: 5 February 2026

Abstract In this study in German higher education, we draw on the first panel of dataset from a longitudinal assessment of domain-specific (DOM) Critical Online Reasoning (COR) in economics in the winter terms 2023/24 (t0) and 2024/25 (t1). DOM-COR tasks were completed as a digital assessment, combining open web search, information evaluation, and reasoning to solve domain-specific scenarios. For this, many students used Artificial Intelligence (AI) chatbots employing Large Language Models (LLMs). This offered a natural quasi-experiment in a real-life learning setting to compare students' DOM-COR performance with and without LLMs. Students also reported on LLM-tools use for study and everyday purposes as well as on their learning outcomes (e.g., credit points, exams). This study investigates both the relationships between LLM-tools use and students DOM-COR assessment performance as well as their learning outcomes at universities. We explicitly distinguish between the qualitative indicators of learning and the speed thereof in the

Dimitri Molerov and Dr. Denis Federiakin share first authorship.

Dimitri Molerov · Denis Federiakin · ✉ Olga Zlatkin-Troitschanskaia · Kevin Shenavai · Lukas Trierweiler · Marie-Theres Nagel
Business and Economics Education, Johannes Gutenberg University, Mainz, Germany
E-Mail: troitschanskaia@uni-mainz.de

Dimitri Molerov
E-Mail: dmolerov@uni-mainz.de

Denis Federiakin
E-Mail: denis.federiakin@uni-mainz.de

Kevin Shenavai
E-Mail: kshenava@uni-mainz.de

Lukas Trierweiler
E-Mail: trierweiler@uni-mainz.de

Marie-Theres Nagel
E-Mail: marie.nagel@uni-mainz.de

context of LLM-chatbots use. By comparing the time students spent completing DOM-COR tasks and their performance between LLM-users and non-users, we examine if, and to what extent, LLM-users had an advantage. While LLM-users were faster, they did not perform better on written task responses. The relationships with learning outcomes trend in the same direction, indicating that regular LLM-users passed more exams, but did not appear to achieve better grades. The findings point to opportunities and risks when using LLM for domain-specific COR-like tasks as well as for regular study assignments.

Keywords Critical Online Reasoning · Digital Assessment · Log Data · Use of AI Tools · ChatGPT · Learning Outcomes

Zusammenhänge zwischen der Nutzung von KI-Chatbots und den Test- sowie Lernergebnissen von Studierenden im Hochschulbereich

Zusammenfassung Diese Studie basiert auf einer längsschnittlichen Untersuchung des kritischen Online-Denkens (Critical Online Reasoning, COR) in der Ökonomie (DOM-COR) im Wintersemester 2023/24 (t0) und 2024/25 (t1). Die DOM-COR Aufgaben wurden im digitalen Assessment bearbeitet und kombinierten offene Websuche, Informationsbewertung und Argumentation, um domänenspezifische Szenarien zu lösen. Mehrere Studierende verwendeten Chatbots mit künstlicher Intelligenz (KI), die Large Language Models (LLMs) einsetzen, um die Aufgaben zu lösen. Diese Situation bot ein ‚natürliches Quasi-Experiment' in den realen Lernsettings, um die Leistung der Studierenden bei DOM-COR Aufgaben ohne und mit Nutzung von LLMs zu vergleichen. Zudem berichteten die Studierenden über die regelmäßige Nutzung von KI-Tools für Studien- und Alltagszwecke sowie ihre Lernergebnisse (z. B. ECTS, Prüfungen). Diese Studie untersucht die Zusammenhänge zwischen der Nutzung von KI-Tools und den DOM-COR-Leistungen der Studierenden *sowie* ihren Lernergebnissen an Universitäten. Dabei unterscheiden wir explizit zwischen den qualitativen Indikatoren des Lernens und der Geschwindigkeit des Lernens im Zusammenhang mit der Verwendung von LLM-Chatbots. Insbesondere durch den Vergleich der Bearbeitungszeit und der Leistung der Studierenden untersuchen wir, ob und inwieweit LLM-Nutzer:innen im Vorteil waren. Während LLM-Nutzer:innen schneller waren, erzielten sie jedoch bei den schriftlichen Aufgabenlösungen keine (signifikant) besseren Ergebnisse. Die Zusammenhänge mit den Lernergebnissen weisen in die gleiche Richtung: LLM-Nutzer:innen absolvierten zwar mehr Prüfungen, erzielten aber offenbar keine besseren Noten. Die Befunde deuten auf Chancen und Risiken hin, die sich aus der Nutzung von LLMs bei COR-ähnlichen Aufgaben sowie im Zusammenhang mit Lernergebnissen abzeichnen.

Schlüsselwörter Kritisches Online-Denken · Digitales Assessment · Log Daten · Nutzung von KI Tools · ChatGPT · Lernergebnisse

1 Research focus and objectives

The rapid development of Artificial Intelligence (AI) has significantly influenced higher education in recent years (Al-Zahrani and Alasmari 2024; Handa et al. 2025). University students report using chatbots such as ChatGPT in their studies (Kelly 2024; Labadze et al. 2023). Given the tendency of Large Language Models (LLMs) to generate hallucinations (Sahoo et al. 2024) and to reproduce unvetted information sources (Zhou et al. 2024; Rejeleene et al. 2024), students are exposed to a flood of unreliable, false, or unchecked information of varying quality, which they must critically evaluate themselves. Owing to the flexibility of LLMs (Herbold et al. 2023; OpenAI 2023), the appeal of LLM-powered chatbots has steadily increased alongside their capabilities (Yu et al. 2024). Although many competing chatbots are available from different providers, ChatGPT remains among the most widely used (Morell-Mengual et al. 2025). In this paper, we therefore do not differentiate between specific chatbots; rather, we generalize findings from research conducted on ChatGPT to LLM-based artificial intelligence chatbots (AI chatbots) more broadly.

The expectations of how AI chatbots influence higher-education students' learning outcomes vary greatly (Wang et al. 2024): while "techno-optimists" claim that LLMs can enhance student learning in unprecedented ways (Kasneci et al. 2023), "techno-pessimists" warn of deskilling, challenges to content authority, and difficulties in assessment delivery (Liang et al. 2023; Royer 2024). Despite the rapid adoption of LLMs, there is a lack of balanced, empirical evidence on their impact on student learning (Bozkurt et al. 2024), particularly regarding Critical Online Reasoning (COR) skills and measurable outcomes such as course credits and exam performance (on student learning outcomes and their measurement in higher education, see Zlatkin-Troitschanskaia et al. 2020). Understanding these effects is crucial for educators and researchers aiming to design effective learning environments, assess digital tool adoption, and maintain rigorous educational standards (Grassini et al. 2024; Swargiary 2024).

A key distinction for assessing learning outcomes in relation to COR lies between Closed Information Space tasks and Open Information Space tasks (Hartig et al. 2025). Closed tasks, such as standardized essays, involve a fixed amount of information with limited access to additional sources (Shavelson et al. 2019). In contrast, Open tasks allow students to draw on virtually unlimited information, constrained only by time, skill, and access (Wineburg et al. 2022). Open tasks more closely reflect real-world problem solving but are rare in educational research, as open-ended search complicates isolating the skill of interest (Mislevy 2018). Nonetheless, assessing LLM use in Open Information contexts is essential to capture their actual impact on authentic problem-solving (Sparks et al. 2024).

Although research on AI tools such as ChatGPT is growing rapidly, many studies are conducted in artificial settings (Wang and Fan 2025). While controlled designs facilitate causal analysis, they often overlook real teaching contexts and learning outcomes such as course credits and exams, limiting practical applicability (Biesta 2007). Investigating the relationships between AI and students' learning outcomes is therefore a necessary step toward effective and scalable integration of these tools in (higher) education (Salas-Martínez and Ramírez-Martinell 2025). Existing research

suggests a highly uneven impact of AI tools: some learning aspects may improve, while others may deteriorate (Grassini 2023).

LLMs' influence on critical thinking, creative thinking, and problem-solving has been a major focus (Zirar 2023; Duenas and Ruiz 2024). Evidence is mixed: some studies suggest AI tools can foster these skills (Gonsalves 2024), while others indicate overreliance or cognitive laziness (Stadler et al. 2024; Putra and Abdunnafi 2024). Differences may stem from definitions of critical thinking (e.g., alertness, reflection, analysis; Oser and Biedermann (2020), or ignoring; Kozyreva et al. 2023), task complexity, and task solvability by chatbots (Drosos et al. 2025; Wang et al. 2024). Despite these nuances, assessment approaches that reflect realistic online conditions remain scarce (Akbar 2025).

To address this gap, we focus on *Critical Online Reasoning* (COR) skills in Open Information Space tasks, combining rigorous assessment with realistic online environments and detailed tracking of student behavior (Schmidt et al. 2020). To observe how students interact with LLM-powered chatbots and to disentangle their effect on real learning outcomes in their academic studies, we analyze domain-specific (DOM) COR tasks and leverage log data (Zlatkin-Troitschanskaia et al. 2025).

In addition to the quality of reasoning, the speed with which students complete complex Open Information Space tasks provides important insights for both educational research and practice (Braun et al. 2020). COR task completion time can reflect students' efficiency in navigating large and unstructured information spaces, their ability to apply critical reasoning strategies under time constraints, and their overall fluency in handling digital information (Leighton et al. 2021; Schmidt et al. 2020). From an educational perspective, understanding these temporal aspects is essential for designing realistic assessments, supporting students in developing strategic information-processing skills, and ensuring that critical reasoning can be effectively demonstrated within limited time frames.

Correspondingly, the study objectives are twofold:

1. To investigate, using log data, the impact of LLM chatbots on students' performance in Open Information Space DOM-COR tasks in economics.
2. To analyze the relationship between students' self-reported habitual use of AI tools for study and everyday purposes and their learning outcomes (e.g., course credits, exams).

Recent research on AI chatbots and learning outcomes highlights a lack of clarity regarding their use in different contexts: the distinction between academic applications and everyday use remains underexplored (Scherr et al. 2025). Habitual everyday interactions with LLMs may themselves contribute to the development of relevant cognitive and metacognitive skills (Stöhr et al. 2024), for instance, through repeated engagement in information search, synthesis, and communication tasks (Sparks et al. 2024). Such informal learning processes may, over time, influence students' reasoning strategies and digital literacy in educational contexts (Zakir et al. 2025). The systematic investigation of these transfer mechanisms represents

a promising direction for future research (Section 5); however, this lies beyond the scope of the present study.

Given students' voluntary use of AI chatbots, we employ quasi-experimental methods to construct a control group. To ensure theoretically grounded matching of control and treatment groups, we leveraged the *Technology Acceptance Model* (Davis 1989). By combining log-data analysis with quasi-experimental design, this study makes a novel contribution. Building on prior research conducted in artificial settings, this study investigates the use of LLM chatbots in authentic learning environments, with a focus on DOM-COR skills and objectively measurable university outcomes. By examining these real-world applications, we aim to address a gap in understanding how LLM chatbots impact academic learning. Our findings provide empirically grounded insights into the opportunities and challenges of integrating LLM-powered chatbots in higher education, addressing a critical gap in the field.

2 Related work

2.1 Critical thinking, learning outcomes and AI chatbots

Recent research has increasingly investigated how students' academic achievements and critical thinking skills change with the use of AI chatbots. Overall, meta-analytical evidence points to predominantly positive short-term associations between chatbot use and learning outcomes (Wu and Yu 2024). Many studies report improvements in academic performance linked to chatbot-supported learning (Mike et al. 2024; Youssef et al. 2024; Adelegan 2024). However, these effects are not uniform and depend on multiple moderating factors (Deng et al. 2025). For example, positive effects are more frequently observed at higher educational levels, whereas long-term effects can diminish over time (Bastani et al. 2024). Some studies even find negative or null effects, particularly when chatbot use replaces rather than supports active learning processes (Swargiary 2024; Abbas et al. 2024).

The present study focuses on critical thinking as a core competence relevant to academic learning. Critical thinking is generally defined as a purposeful, self-regulatory process involving interpretation, analysis, evaluation, and inference, as well as reflective judgment (Facione 1990). Empirically, critical thinking can be assessed through self-reports, standardized tests, or performance-based measures (Braun et al. 2020). Recent research exploring the relationship between AI chatbot use and critical thinking suggests generally positive effects, though findings are heterogeneous due to inconsistent conceptualizations and measurement approaches. Most studies rely on self-reported critical thinking gains, often indicating improvements (Guo and Lee 2023; Youssef et al. 2024; Minh 2024), with some exceptions (Putra and Abdunnafi 2024). A smaller number of studies employing standardized critical thinking measures also support positive associations (de la Puente et al. 2024). However, there remains a marked lack of performance-based assessments of critical thinking in AI-supported learning contexts (Gerlich 2025).

One contributing factor to this gap is that many established critical thinking frameworks were originally designed for static information environments and, as

a result, are inadequately aligned with interactive, adaptive information sources such as LLM-powered chatbots (Gonsalves 2024). These tools can flexibly adjust their responses to user inputs, making them qualitatively different from traditional information sources. Current theoretical developments attempt either to adapt classical frameworks to account for this adaptivity (Moustaghfir and Brigui 2024; Duenas and Ruiz 2024) or to design new frameworks that integrate AI literacy and critical reflection (Knoth et al. 2024; Faust et al. 2023). Yet, appropriate instruments for assessing reasoning skills in such dynamic environments remain scarce, especially in higher education (Wineburg et al. 2022; Nagel et al. 2022, 2024). Moreover, few studies explicitly compare chatbot-supported reasoning in digital assessments with unrestricted Internet access to university learning outcomes, leaving a research gap that this study addresses.

To illustrate the potential impact of chatbots on task performance, consider LLMs such as ChatGPT. These systems can generate a coherent, well-structured essay within seconds, drawing on extensive training data and, in some versions, limited online information. In theory, this could lead to substantial gains in efficiency and response quality with minimal student effort, potentially undermining the validity of assessments if the chatbot effectively completes the task for the student (Tsang and Wood 2025). However, in practice, several challenges often limit these advantages. LLMs may misinterpret task instructions, provide vague or irrelevant content, fabricate sources, or generate outputs that are too lengthy or difficult to process within time constraints. They may also fail to implement requested revisions (Revell et al. 2024). As a result, output quality can vary substantially and cannot be assumed to reliably reflect students' actual competencies. The responsibility for critically evaluating, refining, and integrating chatbot output remains with the student.

Our own baseline pretests with ChatGPT 3.5, conducted prior to the study, corroborated these observations. Although plausible, well-structured responses could be generated quickly, they often lacked topical specificity and were constrained by the model's inability to access current web information at the time. Depending on students' interaction strategies, chatbot use could either support higher-quality reasoning (through time savings that enable reflection, verification, and revision) or lead to uncritical copy-paste behavior, where outputs are accepted without scrutiny, including potential errors or fabricated content. These potential advantages and pitfalls are particularly relevant in open-information critical online reasoning (COR) tasks, where the ability to critically evaluate and integrate AI-generated content can directly affect response quality and completion speed (Federiakin et al. 2024).

2.2 Influences on AI-chatbots use

Much research has investigated individual traits that facilitate or constrain students' use of AI chatbots, as well as the ways in which students engage with these tools (Yu et al. 2024). These studies indicate that students' use of AI chatbots is not random but systematically related to individual characteristics, attitudes, and prior experiences (Morell-Mengual et al. 2025; Stöhr et al. 2024). One prominent theoretical framework in this area is the Technology Acceptance Model (Davis 1989) and its derivatives (Venkatesh et al. 2016), which conceptualize adoption in terms of

perceived ease of use and perceived usefulness, leading to intention to use and actual use. Research on the Technology Acceptance Model has shown that perceived usefulness, ease of use, and social influences shape the adoption of educational technologies, including AI tools (Teo 2011).

For example, Yu et al. (2024) Yu et al. (2024) found significant associations between actual AI chatbot use and both perceived ease of use and perceived usefulness. In contrast, Shaengchart et al. (2024) reported that these factors were not associated with chatbot use when controlling for general attitudes and facilitating conditions. Similarly, Almogren et al. (2024) and Almulla (2024) found that perceived ease of use, but not perceived usefulness, significantly predicted intention to use AI chatbots. Other studies suggest that performance expectancy rather than effort expectancy predicts behavioral intention (Alshammari and Alshammari 2024; Grassini et al. 2024; Romero-Rodríguez et al. 2023). Pan et al. (2024) reported that perceived ease of use moderates the relationship between self-efficacy and intention to use AI chatbots, while perceived usefulness has an indirect moderating effect.

Beyond the Technology Acceptance Model, other covariates have been examined. Abdaljaleel et al. (2024) identified a significant association between AI chatbot use and school graduation. Grassini et al. (2024) and Romero-Rodríguez et al. (2023) highlighted moderating effects of educational and sociodemographic variables on predictors of ChatGPT use. Overall, these variables are often conceptualized as mediators or moderators of AI chatbot use (Deng et al. 2025).

Despite this body of research, findings remain inconclusive and further investigation is needed (Wang and Fan 2025; Heung and Chiu 2025). Methodological choices, such as the use of Partial Least Squares Structural Equation Modeling, which is sensitive to sample size, may partly explain inconsistencies across studies (Hair et al. 2024). Nonetheless, these studies highlight important factors to control for when examining AI chatbot use, which informed our study design (Section 3). Technology Acceptance Model-related variables are particularly relevant for Research Objective 2, as they guide the selection of matching variables to account for self-selection effects in AI chatbot use (Sections 3.3–3.4). This integration clarifies how individual differences in technology acceptance relate to learning outcomes associated with AI chatbot use.

2.3 Critical online reasoning with AI-chatbots

A suitable theoretical framework for studying students' critical thinking skills in highly interactive and adaptive digital environments is *Critical Online Reasoning* (COR; Molerov et al. 2020, Federiakin et al. 2024, Zlatkin-Troitschanskaia et al. 2025). COR conceptualizes critical reasoning as a set of interrelated competencies that students employ when engaging with online information. It comprises three core facets: Online Information Acquisition, Critical Information Evaluation, and Reasoning Based on Evidence and Synthesis (REAS).

Online Information Acquisition refers to students' ability to strategically search for and locate relevant information online. Subfacets include selecting appropriate search engines, databases, or platforms, narrowing and specifying search problems, formulating effective search terms, navigating complex websites, and excluding low-

quality sources when better alternatives are available. Performance indicators comprise navigation processes and the quality of sources identified.

Critical Information Evaluation involves assessing the reliability and validity of online information. For instance, students are expected to attend to credibility cues such as author expertise, transparency of sources, or institutional affiliations, and to distinguish between actual and merely claimed quality indicators (e.g., invented quality seals). Performance is reflected in students' ability to provide justified, contextually relevant criteria for their source selections.

REAS covers the synthesis and integration of information from multiple sources into coherent, evidence-based arguments. Subfacets include the coherence and rigor of reasoning, the number and quality of arguments, the use of relevant evidence, and the consideration of multiple perspectives. Performance is assessed on the basis of the quality of the students' written responses (short essays).

The COR framework builds on two important predecessors. First, it extends the Information Problem-Solving on the Internet model (Brand-Gruwel et al. 2009), which describes phases of online information seeking and problem solving, but is less explicit regarding evaluation quality criteria and the REAS facet. Second, it aligns with the open-web search assessment approach (Wineburg et al. 2022), which allows students unrestricted Internet access to locate or verify information. This approach requires strategic, critical engagement with a wide range of online materials. COR integrates these perspectives by combining a process view (how students search, evaluate, and reason) with a performance view (the quality of their resulting products) (Schmidt et al. 2020).

Importantly, COR distinguishes between generic application contexts and domain-specific contexts (Nagel et al. 2024). Domain-specific COR tasks are embedded within disciplinary knowledge structures and practices. In the present study, we focus on the domain of economics, as it exemplifies contexts in which online information seeking, evaluation, and synthesis are integral to both academic learning and professional practice. Economic tasks frequently require integrating diverse and sometimes conflicting information from multiple sources, constructing complex arguments, and applying evidence-based reasoning; i.e., processes that align closely with the COR framework. Economics education emphasizes critical evaluation, argumentation, and engagement with real-world information environments, making it particularly suitable for examining domain-specific COR. By focusing on domain-specific contexts, we can, first, examine how students apply critical reasoning in authentic, content-rich tasks—yielding insights that would remain less evident in more generic settings—and, second, link these measures to actual learning outcomes in higher education.

Domain-specificity is further conceptualized using the *Model of Domain Learning* (Alexander 2004), which distinguishes three developmental stages of learning: Acclimation, Competence, and Proficiency. This framework provides a theoretical basis for structuring and interpreting domain-specific (DOM) COR tasks. We expanded this part to clarify that the Model of Domain Learning framework informs both the operationalization of DOM-COR tasks and the rationale for their relevance in economic education, where navigating complex, information-rich environments is a core competency. In line with the Model of Domain Learning, we differen-

tiate between three DOM-COR application contexts: *Fundamental, Practical, and Transdisciplinary*. This structuring allows us to capture varying levels of task complexity and domain integration, thereby enabling a more precise analysis of students' reasoning processes within the economics domain.

With the emergence of LLM-based chatbots, the COR framework has recently been linked to the framework of Prompt Engineering skills (Federiakin et al. 2024). While LLMs can autonomously perform several substeps of the three COR facets, their outputs remain machine suggestions that require critical scrutiny. For example, chatbots can search (internally), retrieve, and synthesize information, but they may also generate fabricated content, misinterpret prompts, or fail to provide source transparency. This introduces new challenges for students: They must decide whether to trust, refine, or reject chatbot outputs and, if necessary, reformulate prompts or consult additional sources.

Crucially, information retrieval with AI chatbots differs from conventional web search in both quantitative and qualitative respects. Quantitatively, chatbots can substantially reduce the time required to generate initial drafts, arguments, or summaries. This shift has the potential to reallocate students' cognitive resources from information retrieval toward critical evaluation and synthesis. Qualitatively, AI chatbots frequently deliver pre-structured but often opaque outputs, making it more difficult for students to verify the provenance and credibility of the underlying information. Although a comprehensive theoretical and empirical analysis of these qualitative differences lies beyond the scope of a single study, these developments underscore the continued and even increasing relevance of the COR framework for understanding student competencies in AI-augmented learning environments (Federiakin et al. 2024). COR provides a theoretically grounded lens for analyzing how students search for information, critically evaluate AI-generated content, and integrate chatbot output into their reasoning processes. In sum, COR constitutes a suitable framework for examining how AI chatbot use may influence both the quality of students' reasoning and their task efficiency in realistic, domain-specific problem-solving contexts.

2.4 Research questions

Existing research on the impact of AI-chatbots on student learning outcomes and critical thinking skills has several limitations (Schei et al. 2024). Many studies have been conducted under highly controlled and standardized conditions or rely on self-reported measures (Wang and Fan 2025), which restricts their ecological validity and generalizability to authentic learning contexts. Moreover, as outlined in Sections 2.1–2.3, findings remain heterogeneous and sometimes contradictory, reflecting differences in methodological approaches, task designs, and definitions of critical thinking and reasoning skills.

Against this background, our study addresses two central research objectives (Section 1). To address Research Objective 1, we compare the performance of students who used AI chatbots during the DOM-COR assessment with those who did not. As the decision to use or not use chatbots during the assessment represents a non-random self-selection process influenced by personal and contextual factors (Liang et al. 2023), we applied a quasi-experimental matching technique to cre-

ate comparable groups and to control for potential confounding variables. In line with the Technology Acceptance Model (Davis 1989), variables related to perceived usefulness, ease of use, and prior experience with AI tools were used as matching variables. This integration of the Technology Acceptance Model is theoretically motivated: Regular use of AI chatbots for everyday purposes may shape students' familiarity and attitudes toward the technology, which in turn can influence their adoption in academic contexts (Research Objective 2).

To address Research Objective 2, we examine associations between students' self-reported habitual AI chatbot use and their university learning outcomes. The boundaries between everyday and academic use of AI chatbots are inherently fluid. Everyday interactions with LLMs may foster skills relevant to academic contexts, such as information searching, synthesis, and critical evaluation, albeit in less structured forms. Such informal learning experiences could influence students' academic performance through skill transfer, yet empirical research on this transfer remains scarce (Sect. 1). Consequently, our study treats students' chatbot use during DOM-COR assessments and their habitual use patterns as complementary but distinct indicators of engagement with AI tools.

To operationalize student performance in the DOM-COR assessment, we use two indicators:

- a qualitative indicator (the REAS subtask score), which reflects the quality of students' reasoning and synthesis based on evidence, and
- a quantitative indicator (task completion time) which captures task processing speed (Wang et al. 2023).

By examining these two indicators, we address a critical gap in the literature: previous studies have often focused exclusively on performance quality or efficiency measures, but rarely on both simultaneously (Kaiss et al. 2024). Distinguishing between speed and quality is particularly relevant in AI-augmented learning environments, where chatbots can accelerate certain task phases (e.g., drafting) without necessarily enhancing the quality of reasoning. This distinction further implies that improvements in task completion speed may manifest immediately, whereas gains in reasoning quality depend on students' capacity to critically engage with, refine, and validate chatbot outputs.

Based on the current state of research, the theoretical grounding in COR and the Technology Acceptance Model, and the identified research gaps (Sect. 2.1–2.3), we formulate the following research questions (RQs):

1. Do students who use AI-chatbots demonstrate higher-quality, evidence-based, and synthesis-based reasoning when completing DOM-COR tasks?
2. Are students who use AI-chatbots significantly faster at completing DOM-COR tasks compared to non-users?
3. Are there significant associations between students' habitual AI chatbot use and their university learning outcomes (considering both quality and quantity indicators)?

Specifically, RQ 1 and RQ 2 address the direct impact of chatbot use during complex reasoning tasks (DOM-COR), whereas RQ 3 examines broader patterns related to everyday AI tool use and university learning outcomes. Together, they enable a differentiated analysis of the role of AI-chatbots in higher education.

3 Methods and data

3.1 Study design

The data were collected as part of an ongoing longitudinal project (Critical Online Reasoning in Higher Education (CORE)) aimed at tracking the development of students' Critical Online Reasoning (COR) skills and identifying factors that promote these skills over the course of bachelor's studies. The first measurement occasion took place in the winter term 2023/24 (t0), the second in the winter term 2024/25 (t1).

At t0, a subset of students used AI chatbots while performing the DOM-COR tasks on the Internet, whereas the majority of participants did not. This situation created a natural experiment, allowing a comparison of task performance between self-selected users and non-users.

At t1, to keep assessment conditions as fair as possible, access to all known and identified AI chatbots was prohibited. Hence, the available performance and timing data on the identified AI-chatbot users are cross-sectional and from t0 (between-subjects design). Regarding further development, upon repeated participation at t1, the data collected from students included self-reported learning outcomes after one year of studies (within-subjects design).

3.2 Sample

Participants for the longitudinal COR study were recruited through a complete survey of all incoming students in the participating study programs, namely, economics and economics teacher education, with some students majoring in the social sciences, at five German universities during introductory lectures.

Table 1 DOM-COR sample descriptive statistics

Variable	Minimum	Maximum	Mean	SD
Gender (1 = Female)	0	1	0.53	N/A
Age	15	57	20.17	3.025
High School GPA[a]	0.9	3.5	1.74	0.606
No prior university studies	0	1	0.80	N/A
No prior apprenticeship	0	1	0.85	N/A
Semester	1	1	1.00	0.000
DOM-COR (1 = EC)	0	1	0.57	N/A

Table 2 The number of students after data cleaning who used or did not use LLM-tools across the four DOM-COR tasks

DOM-COR Task	Used LLMs	Did not use LLMs	Total
Nudging	16	69	85
Pilot strike	16	70	86
Startup	17	74	91
Wind Farm	6	78	84

3.2.1 Quasi-experimental analysis sample (RQ1 & RQ2)

For this paper, we used data from 270 first-semester students majoring in the social sciences ($N = 97$) and economics, including teacher education ($N = 173$) who met the following criteria: they (1) responded to DOM-COR tasks in economics (Table 1), (2) completed the corresponding survey on habitual AI-chatbots use, and (3) reported their university learning outcomes.

AI-chatbot user identification and verification in the DOM-COR assessment log data were based on detected access to AI-chatbots. Candidate cases were initially identified through a keyword search (e.g., ".ai"), after which the log events of all included students were manually verified for indicators of AI-chatbots use. Any necessary reclassifications were performed iteratively until all misclassifications were resolved.

During data cleaning, 25 observations (out of 270) with missing values in at least one of the selected variables (Section 3.4.2.2) were removed, as their inclusion would have hindered the matching procedure (Section 3.4). Of the remaining 245 observations, 38 students had used AI chatbots, whereas 207 had not (Table 2). The final sample had a mean age of 20.17 years (SD = 3.03) and was 53% female.

3.2.2 Cross-sectional analysis sample (RQ3)

To investigate RQ3, we used data from the repeated observation of the same students at t1, i.e., after one year of their bachelor's studies. Given an attrition rate of approximately 36%, $N = 171$ (mean age = 20.40 (age SD = 2.51), 57.3% females) of the students who took part at t0 remained at t1.

3.3 Materials and instruments

3.3.1 DOM-COR assessment, tasks, and scoring rubrics

COR skills are assessed using recently developed and tested COR assessments (Nagel et al. 2022, 2024; Hartig et al. 2025). The assessment comprises several COR tasks administered at random (Table A1, Appendix).

In this paper, we used four DOM-COR tasks in economics. Each task was designed in three variations corresponding to the fundamental, practical, and transdisciplinary application contexts as defined based on the Model of Domain Learning (Alexander 2004) (Section 2.3). Each task presents a scenario-based information problem without a single correct solution. Students were encouraged to use the In-

ternet and search for relevant information via search engines of their choice. Based on the search and critical evaluation of selected information, students compile a short essay as the task response.

During the assessment, students had 60 min per DOM-COR task with two application contexts. Two-thirds of participants received two DOM-COR tasks, while one-third received a single task at random. Each task was administered in two of the three application contexts, and the order of tasks and contexts was randomized. Students completed the assessment on a virtual machine from home. All actions, including task times, text entries, web searches, content viewed, and timestamps, were logged and stored for analysis. These log data supplemented the scoring of the students' written responses to the REAS subtask and were used to analyze task completion time. Participation was self-paced, allowing students to choose when to take the test and potentially pause between tasks within the two-month assessment window. Participants were compensated 50 euros for completing the full assessment.

The scoring rubrics for the written responses were standardized across all three contexts (Table A1, A2 Appendix). Rubrics focused on the quality of the written response, including criteria such as addressing the problem, coherence, stringent argumentation, number of arguments from the domain, balance, conclusion provided, and perspectives from other domains (in the transdisciplinary context). A detailed description of the scoring scheme is provided in Hartig et al. (2025).

Each response was scored independently by at least three trained raters. Interrater reliability was assessed using Cohen's kappa (Sim and Wright 2005) for each rater pair across all REAS criteria within each task. Summary results by task are presented in Table 3. The overall reliability across tasks was high, with an intraclass correlation coefficient (ICC) of 0.774 (Liljequist et al. 2019), indicating substantial agreement among raters. This supports the use of averaged scores across raters and criteria for subsequent factor analyses (Cronbach and Gleser 1964).

As a DOM-COR measure, we used maximum a posteriori factor scores (Skrondal and Rabe-Hesketh 2004) derived from a factor-analytic model. The reliability of the general factor score was estimated at 0.760 using the composite reliability index (Bentler 1968; often referred to as McDonald's (1970) ω), suggesting its suitability for group-level analyses.

Table 3 Interrater agreement per task

Interrater agreement	Percent of rater pairs per task			
	Pilot strike (%)	Nudging (%)	Windfarm (%)	Startup (%)
Slight (<0.2)	8.3	11.1	1.4	9.7
(0.2<) Fair (<0.4)	37.5	23.6	22.2	25.0
(0.4<) Moderate (<0.6)	36.1	30.6	37.5	27.8
(0.6<) Substantial (<0.8)	9.7	13.9	18.1	19.4
Almost Perfect (>0.8)	8.3	20.8	20.8	18.1

Cohen's κ is defined for a pair of raters who score the same items. For each task, we computed κ for every possible pair of raters on every rating criterion. With R raters this yields R(R − 1) / 2 rater pairs per criterion (e.g., with three raters there are three pairs). The resulting κ values were then grouped into the agreement categories shown in the table rows (Slight, Fair, etc.). For each task, the table reports the percentage of rater pairs—aggregated across all criteria—whose κ falls into each category

3.3.2 Log data: time on task

Students were allotted 60 min per DOM-COR economics task, which included two application contexts. A visible timer tracked time on task, and start and submission events were automatically time-stamped. For the analysis of response time, we used the total time taken to complete the DOM-COR task, summing the time spent on both contexts within a task. This sum was treated as the total task completion time. Following recommendations from the response time literature (Lo and Andrews 2015), we did not transform this variable (e.g., by taking its logarithm), ensuring interpretability in subsequent analyses.

3.3.3 Log data: AI-chatbots use in assessment

Students' access to AI chatbots during the DOM-COR assessment was determined by analyzing the log data, which recorded all navigation and input events, including website URLs visited and content typed or copied. Candidate cases were identified using keywords such as ".ai," "GPT," and the most commonly used AI chatbots at the time, with additional instances identified through snowballing. Approximately two-thirds of logged AI tool uses were OpenAI's ChatGPT. Other AI-chatbots or search functions used included Microsoft Bing, Google Gemini, and occasional instances of less common tools, such as Poe (poe.ai).

3.3.4 Self-reported habitual AI use

In addition to the DOM-COR assessment, students completed a comprehensive contextual survey. Beyond sociodemographic information, the questionnaire assessed students' media use both in everyday life ("How often do you use the following media on average per week to obtain information in your daily life?") and for learning purposes ("How often do you use the following media on average per week when working on assignments or searching for information for your studies?"). Each item referred to a specific media source, with AI-chatbots included as a separate category. Responses were assessed on a 6-point Likert scale, ranging from 1 (never) to 6 (multiple times daily). These measures were used to construct variables reflecting students' habitual AI-chatbots use in both everyday and university contexts.

3.3.5 Self-reported competence in the media and AI use and their relevance for studies

By t1, the context survey was slightly revised to include additional questions on students' self-assessed competence ("Please rate your own skills in comparison to those of other Internet users.") in using the media sources and the perceived relevance of these skills for their studies ("Please rate the importance of these skills for your studies.").

Each item was assessed for all media sources, with AI-chatbots included as a separate category. Responses were assessed on a 5-point Likert scale, ranging from 1 (very low) to 5 (very high). These measures capture students' perceived

ability to use different media effectively and the importance they attribute to these skills in academic contexts.

3.3.6 University entrance qualification

At t0, the context survey also included students' final school grades (German Abitur) overall and in English, German, and Mathematics (Table A3, Appendix). As a generalized measure of university entrance qualification, we summarized these final overall school grades using principal component analysis. The first principal component explained 70.65% of the variance in the observed variables (the second explained 16.91% and the third 12.44%). Hence, we used the first principal component as an indicator of general academic ability at university entry (for the factor loadings of the variables on the first component, see Table A4, Appendix).

3.3.7 Student learning outcomes

At t1, key student learning outcomes were collected via self-reports, including the number of courses attended and passed, grades obtained, and total credit points accumulated. Given that learning outcomes cannot be meaningfully captured through single, isolated indicators (Zlatkin-Troitschanskaia et al. 2020), we drew on a set of complementary measures to provide a more comprehensive account of students' academic progress, encompassing course participation and completion, performance outcomes, credit accumulation, and exam success throughout their university studies. For this, we used the number of courses students had taken between t0 and t1, categorized as "0," "1," or "2 or more," and the average grades from these courses. We derived interaction terms between these variables for corresponding courses and used principal component analysis to compress these variables into one generalized learning outcomes index. Since none of the variables were standardized, the interaction variables reflect the expected sum of the grade points students gained from all the courses they took. The first principal component from the set of interaction variables explained 33.95% of the variance (the second 17.07% and the third 13.46%), with a continued linear decrease and eigenvalues less than 1 (for the factor loadings on the first component, see Table A5, Appendix).

3.3.8 Psychological constructs

Additionally, the survey collected information on several psychological constructs reported to be relevant for media use, including the use of AI chatbots. Specifically, we assessed Internet-specific epistemic beliefs (Chiu et al. 2013) consisting of four subscales: Justification for Internet-based knowledge (Cronbach's $\alpha = 0.544$), Internet-based knowledge sources ($\alpha = 0.688$), Simplicity of Internet-based knowledge ($\alpha = 0.653$), and Certainty about Internet-based knowledge ($\alpha = 0.623$). Also, we assessed unidimensional information-behavior-related self-efficacy expectations (Behm 2018) ($\alpha = 0.824$). These raw scores were used in the quasi-experimental analyses.

Finally, we controlled for students' tendencies to use systematic and heuristic information processing, as these scales have been shown to be relevant in predicting student learning outcomes (Schemer et al. 2008). We used latent variable modeling consisting of two scales: one reflecting the tendency to use a heuristic (quick, holistic, and intuitive) way to process information ($\alpha = 0.745$), and one measuring the tendency to rely on a systematic (slow, argumentative, and rigorous) way of processing information ($\alpha = 0.713$).

3.4 Data analysis

3.4.1 Quasi-Experimental analysis (RQ1 & RQ2)

Quasi-experimental analysis (Kim and Steiner 2016) allows researchers to isolate the average treatment effect by non-randomly subsampling observations from the "control" group so that the observed distributions of the covariates (balancing variables) across the two groups are equivalent. This helps minimize the likelihood of confounding the causal claim by reducing differences between the two groups in the observed covariates. Specifically, we used a genetic algorithm (Diamond and Sekhon 2013) to optimize the Generalized Mahalanobis Distance between the two groups. For the analysis, we used the MatchIt package v. 4.5.5 (Ho et al. 2023) in R v. 4.4.1.

As some students used AI-chatbots for only one of two administered DOM-COR tasks, we conducted matching within each task separately. A few students appeared twice in the sample: they used AI-chatbots in one task and did not use them in the other. The unit of analysis was therefore a student-by-task interaction.

3.4.1.1 Matching and balancing variables To select appropriate matching variables, we analyzed the literature to identify factors influencing students' use of AI-tools (Section 2.2). Several variables were used as proxies for the components of the Technology Acceptance Model. Specifically habitual media use in everyday life served as a proxy for perceived ease of AI use, assuming that frequent use of certain media reflects ease of use. Additionally, the information-behavior-related self-efficacy expectation scale was employed as a complementary proxy, capturing how easily students can use the Internet to solve various information problems.

As a proxy for the perceived usefulness, we used measures of Internet-specific epistemic beliefs, assuming that these reflect students' perceptions of the utility of AI for solving information problems. Finally, habitual media use for learning purposes served as a proxy for intention to use AI, reflecting the transfer of perceived usefulness to domain-specific tasks.

To further ensure comparability between groups, we balanced for students' study major, since social sciences and economics students completed the same DOM-COR assessment, yet may employ different response strategies. Economics students, for instance, might rely more heavily on domain-related knowledge. Additional balancing variables included school graduation in German, English, and Mathematics, as well as overall grade, alongside gender, age, socioeconomic, and cultural-linguis-

tic background (Yudytska and Androutsopoulos in press; Abdaljaleel et al. 2024; Grassini et al. 2024; Romero-Rodríguez et al. 2023).

In total, 48 observed variables were used to achieve balanced matching between students who did and did not use AI-chatbots during the assessment.

3.4.1.2 Imbalance analysis in the matched groups After matching an equivalent observation from the non-user group to each observation from the AI-chatbot user group, we analyzed imbalance in the observed variables across groups (Table A3, Appendix). Although some covariates were moderately imbalanced, there are no definitive cutoffs; in practice, an average Kolmogorov-Smirnov statistic of 0.10 or lower (Stuart et al. 2013) is often considered to indicate acceptably balanced groups. On average, the matched groups can be treated as comparable, with an average standardized mean difference of 0.03 and an average Kolmogorov-Smirnov statistic of 0.12.

3.4.1.3 Groups comparisons For group comparisons, we used regression models, which generalize t-tests and allow control for several variables. For the analysis of task competition times, we employed mixed-effects models estimated via restricted maximum likelihood (using lme4 package v. 1.1-37 for R), with Satterthwaite's method for determining the statistical significance of t-tests (Kuznetsova et al. 2017; lmerTest package v. 3.1-3 for R). Mixed-effects models were chosen since matching was conducted within each task and, due to the assessment design, most participants were given two tasks (Table A1, Appendix).

For the analysis of student task performance, we used a simple linear model (a base function of R v. 4.5.0) with dummy variables indicating the tasks participants received and whether they used AI chatbots while solving the task. This approach was chosen as some participants were given two tasks, but only a single general factor score was used for each student. Hence, controlling for the assigned task via dummy variables yields a full-rank predictor matrix.

3.4.2 Cross-sectional analysis (RQ3)

3.4.2.1 AI-chatbots use and control variables Since the list of questions at t1 was slightly modified compared to t0, we adjusted the proxies used in the cross-sectional analysis relative to the quasi-experimental analysis. Self-assessed competence in using AI-chatbots served as the proxy for perceived ease of use, based on the well-established positive relationship between mastery of a task and perceived ease in executing it (Rodgers et al. 2014).

The perceived usefulness of AI-chatbots and the actual use of AI-chatbots for learning purposes were self-reported by students, measured respectively as the perceived relevance of AI-chatbot-related skills for their studies and the frequency of AI-chatbots use for study purposes.

In all Technology Acceptance Model-related mediation schemes, we additionally controlled for habitual AI-chatbots use in everyday life, assuming that this variable captures students' general propensity to use AI-chatbots across contexts.

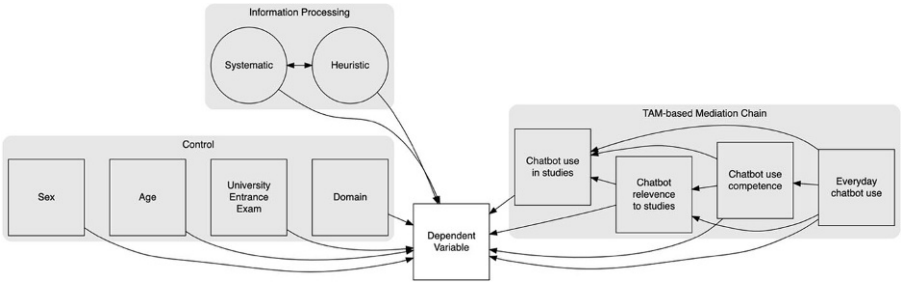

Fig. 1 The general path diagram of the mediation models

3.4.2.2 Mediation model structure For the within-subject cross-sectional analysis, we specified a model based on a chain of Technology Acceptance Model-derived mediations. We controlled for sociodemographic variables (age and sex), study domain (economics and social sciences), and university entrance qualification as an indicator of pre-tertiary learning outcomes.

The mediation model examines the influence of competence in using AI-chatbots on student learning outcomes via a chain of mediators, namely, the perceived relevance of AI-chatbots skills for studies and the actual use of AI-chatbots for learning purposes. In this mediation chain, we also controlled for habitual AI-chatbots use in everyday life, serving as a proxy for students' general tendency to use AI-chatbots for solving information problems.

Row-wise deletion of missing data was applied, and the model was estimated using the lavaan package v. 0.6–19 for R. The structure of the mediation model used in the correlational analyses is illustrated in Fig. 1.

4 Results

4.1 COR performance (RQ1)

The results of the group comparison regarding students' task performance are presented in Table 4.

The residual standard deviation was 0.607 (df = 68). As shown in Table 6, students' task performance did not differ significantly between AI-chatbot users and

Table 4 Comparison of factor scores

Coefficient	Value	S.E.	t-value	*p*-value
Intercept	0.012	0.222	0.053	0.958
LLM use (1 = Y, 0 = N)	0.287	0.163	1.761	0.083
Nudging	0.234	0.214	1.096	0.277
Pilot strike	–0.011	0.213	–0.052	0.959
Startup	0.286	0.225	1.269	0.209
Windfarm	0.273	0.233	1.169	0.247

Table 5 Response time analysis

Fixed effects					
Coefficient	*Value*	*S.E.*	*Df*	*t-value*	*p-value*
LLM use (1 = Y, 0 = N)	–476.69	213.64	71.14	–2.231	0.029
Nudging	2385.36	177.23	85.76	13.460	<0.001
Pilot strike	2479.93	169.56	88.11	14.625	<0.001
Startup	2400.07	169.25	88.75	14.181	<0.001
Windfarm	2821.00	215.52	66.80	13.089	<0.001
Random effects					
Variable	*SD*				
ID	833.4				
Residual	394.0				

non-users for any of the tasks. Users exhibited a slight advantage over non-users, but this difference was not statistically significant. Overall, comparisons of factor scores indicated no statistically significant differences in students' DOM-COR performance between the AI-chatbot and non-chatbot conditions.

4.2 Task completion time (RQ2)

Results of the response-time analysis are presented in Table 5.

As shown in Table 5, response times differed significantly between AI-chatbots users and non-users across all tasks. Therefore, participants who used AI-chatbots completed the DOM-COR tasks more quickly than non-users.

4.3 Relations with student learning outcomes (RQ3)

Results concerning students' learning outcomes are summarized in Table 6.

Since the Technology Acceptance Model-based chain of mediations remained consistent across three models, the corresponding coefficients are presented separately in Table 7.

The regression results indicate that the models fit the data adequately. AI-chatbots use in everyday life was positively associated with the number of exams taken and passed, but it did not show a significant relationship with aggregated course grades. Importantly, no other variable in the Technology Acceptance Model-based mediation chain was significantly associated with student learning outcomes, including the number of exams taken or passed. Additionally, the heuristic information-processing scale score exhibited a marginally negative association with the number of exams passed after the first year of studies. Regarding sociodemographic control variables, University Entrance Qualification was positively related to student learning outcomes, and sociology students tended to take fewer exams while achieving higher grades than economics students.

Within the Technology Acceptance Model-based chain of mediations, all paths were statistically significant, providing support for the theoretical derivations.

Table 6 Relations with students' learning outcomes (SLOs)

Independent variables		Dependent variables								
		SLOs			Exams taken			Exams passed		
		Est. (S.E.)	*p*-value	Std. Est.	Est. (S.E.)	*p*-value	Std. Est.	Est. (S.E.)	*p*-value	Std. Est.
Information processing	Systematic	0.092 (0.216)	0.671	0.048	–0.566 (0.333)	0.089	–0.217	–0.575 (0.370)	0.120	–0.209
	Heuristic	–0.203 (0.201)	0.311	–0.106	–0.611 (0.334)	0.067	–0.234	–0.762 (0.384)	0.047	–0.278
AI-chatbots	Competence	0.051 (0.185)	0.783	0.031	–0.113 (0.213)	0.597	–0.051	–0.292 (0.197)	0.137	–0.125
	Relevance	–0.106 (0.156)	0.496	–0.061	–0.011 (0.219)	0.962	–0.004	0.100 (0.217)	0.644	0.041
	Use in studies	–0.020 (0.135)	0.882	–0.017	–0.308 (0.212)	0.147	–0.188	–0.260 (0.226)	0.249	–0.151
	Everyday use	0.201 (0.195)	0.302	0.150	0.476 (0.220)	0.030	0.261	0.533 (0.216)	0.014	0.277
Control variables	University Entrance Qualification	0.291 (0.084)	0.001	0.231	–0.037 (0.137)	0.788	–0.021	0.047 (0.142)	0.743	0.026
	Gender	0.144 (0.253)	0.571	0.037	–0.368 (0.399)	0.357	–0.070	0.220 (0.420)	0.600	0.040
	Age	0.038 (0.068)	0.579	0.049	–0.043 (0.088)	0.630	–0.041	–0.038 (0.102)	0.711	–0.034
	Domain: Sociology	0.863 (0.275)	0.002	0.212	–1.721 (0.503)	0.001	–0.310	–1.031 (0.534)	0.054	–0.177
Model Fit										
Baseline Model	χ^2	721.841			721.436			715.762		
	d.f. (χ^2)	126			126			126		
	p-value	0.000			0.000			0.000		
	Scaling factor	1.013			1.053			1.054		

Table 6 (Continued)

Independent variables		Dependent variables								
		SLOs			Exams taken			Exams passed		
		Est. (S.E.)	*p*-value	Std. Est.	Est. (S.E.)	*p*-value	Std. Est.	Est. (S.E.)	*p*-value	Std. Est.
Tested model	χ^2	160.744			159.728			159.118		
	d.f. (χ^2)	100			100			100		
	p-value	0.000			0.000			0.000		
	Scaling factor	1.013			1.012			1.015		
	SRMR	0.065			0.065			0.064		
	Robust CFI	0.907			0.907			0.907		
	Robust TLI	0.882			0.883			0.883		
	Robust RMSEA	Estimate	0.059			0.058			0.058	
		90% CI	(0.040, 0.076)			(0.040, 0.075)			(0.040, 0.075)	

Table 7 The coefficients from the Technology Acceptance Model-based chain of mediations

Independent variables	Effects		
	Est. (S.E.)	*p*-value	Std. Est.
Dependent variable: AI-chatbots competence			
AI chatbots in everyday use	0.379 (0.065)	<0.001	0.463
Dependent variable: AI-chatbots relevance			
AI-chatbots competence	0.285 (0.085)	0.001	0.301
AI-chatbots in everyday use	0.172 (0.066)	0.010	0.223
Dependent variable: AI-chatbots use in studies			
AI-chatbots competence	0.192 (0.068)	0.005	0.141
AI-chatbots relevance	0.184 (0.07)	0.009	0.128
AI-chatbots in everyday use	0.762 (0.053)	<0.001	0.684

5 Conclusion

5.1 Discussion

Expectations regarding the educational potential of AI-chatbots are steadily increasing, yet the problem of hallucinations remains unresolved (Kalai et al. 2025). Consequently, students using such tools—by now the majority—are exposed to unverified, potentially misleading, or factually incorrect information. This raises important questions concerning how students critically engage with and reason from information retrieved via AI-chatbots. A key issue in this context is the impact of chatbot use on students' task performance and learning outcomes. While some studies report positive associations between chatbot use and student learning outcomes (Wang and Fan 2025), others highlight predominately negative or null effects (Kosmyna et al. 2025). Such discrepancies are often attributable to methodological and sampling differences, underscoring the need for quasi-experimental studies with larger samples and real-world educational outcome measures (Laun and Wolff 2025).

The present study investigated the relationship between chatbot use and student performance on authentic digital assessments of Critical Online Reasoning (COR) in economics (response quality—RQ1; response times—RQ2) as well as real-world student learning outcomes (RQ3). To address RQ1 and RQ2, a quasi-experimental design was applied to control for self-selection effects in chatbot use. RQ3 was examined using regression analyses to estimate associations between chatbot use intensity and indicators of study productivity during the first academic year.

With regard to task processing speed, chatbot users completed DOM-COR tasks significantly faster than non-users (RQ2). In contrast, group comparisons of factor scores revealed no statistically significant differences in task performance (RQ1). Although chatbot users achieved slightly higher factor scores on average, this difference was not statistically significant. This indicates that chatbots with capability comparable to ChatGPT 3.5 supported students in working more efficiently but not in achieving higher quality task responses. Importantly, these findings are based on group comparisons rather than pre-post analyses and therefore do not allow causal conclusions regarding performance improvements (Section 5.2).

The correlational analyses conducted for RQ3 yielded a complementary pattern, while not permitting causal inference. Students who reported frequent everyday chatbot use took and passed more examinations during their first year of study compared to those who used chatbots less frequently. However, chatbot use was not significantly related to qualitative indicators of learning, such as course grades. This suggests that chatbot use may increase the quantity of academic achievements (e.g., number of exams passed), without affecting their quality. Similar conclusions have been drawn in recent studies reporting limited effects of AI chatbot use on productivity and learning quality in higher education (Mohamed et al. 2025; Challapally et al. 2025).

These findings have important conceptual implications. The observed pattern suggests that chatbots may enable students to increase throughput—the number of tasks completed per unit time—without necessarily enhancing cognitive depth or learning quality. When chatbots are used beyond students' cognitive capabilities, users may lose control over output quality, which can negatively affect reasoning processes and task performance. This interpretation is consistent with recent work indicating that AI-tools may accelerate task execution while potentially undermining deeper learning processes if critical engagement is insufficient (Kosmyna et al. 2025; Deng et al. 2025).

Notably, when everyday chatbot use was statistically controlled, study-specific chatbot use variables no longer exhibited significant associations with learning outcomes. This contrasts with meta-analytic evidence showing positive effects of chatbot use on learning outcomes (Deng et al. 2025), though such studies often rely on short-term interventions in less realistic settings. At the same time, variables derived from the Technology Acceptance Model remained significantly interrelated, supporting earlier research on technology acceptance in economics and social sciences (Yu et al. 2024; Strzelecki 2024). The pattern indicates that domain-specific AI-chatbot use in higher education may not yield additional productivity benefits beyond those associated with everyday chatbot use.

5.2 Limitations

When interpreting these findings, several limitations must be considered. First, the absence of significant performance differences between AI-chatbot users and non-users may partly reflect limited statistical power, due to the relatively small subgroup of chatbot users. Moreover, heterogeneity in the extent and manner of chatbot use—ranging from extensively use across all tasks to minimal use for single subtasks—may have increased within-group variability and attenuated between-group differences.

Second, the analyses focused on tasks from a single academic domain (economics) and did not fully control for potential study-domain differences, such es initial familiarity with chatbots or domain-specific epistemic beliefs. These factors may act as moderators of chatbot effects.

Third, the study focused mainly on product data, while process data (e.g., detailed log files) were only partially analyzed. This limits our ability to draw conclusions about the mechanisms underlying students' interactions with AI-chatbots during

problem solving. Given that chatbots are expected to influence information problem-solving processes, this represents a substantive limitation. Future research could investigate different AI-chatbots use strategies or levels of proficiency (Federiakin et al. 2024; Knoth et al. 2024) as potential moderators. For instance, analyzing differences in outcomes across various strategies of AI-chatbots use may help identify both effective and inadvisable approaches to employing chatbots in information problem-solving tasks. Likewise, varying levels of proficiency in AI-chatbots use could moderate the impact of chatbots on students' actual performance.

Fourth, this study—as with most research in this field—is "historically" situated, reflecting both the technological capabilities of ChatGPT 3.5 and students' usage habits at a specific point in time. For instance, while ChatGPT 3.5 could not perform Internet searches at the time of the study, current standard LLMs can search the web, filter content in relation to a prompt, and generate reasoning chains. Such rapid technological developments, including integrated web search and advanced reasoning capabilities, may alter students' search behaviors, familiarity with tools, and critical engagement, thereby complicating replication and longitudinal measurement.

Finally, correlational analyses do not allow for causal conclusions, and quasi-experimental comparisons cannot substitute for randomized controlled trials. Interpretations have therefore been carefully framed in terms of associations and group differences.

5.3 Implications

Major strengths and weaknesses of AI-chatbots use for education, including higher education, have been identified (Farrokhnia et al. 2024): Highlighted strengths have been discussed in regard to efficiency, such as generation of plausible answers, self-improving capability, increased access to information, personalized real-time responses and complex learning, lower teaching workload; in turn, suspected weaknesses include AI-chatbots use promoting a lack of deep understanding and activation of higher-order thinking skills, difficulty in evaluating response quality and overreliance on responses, along with risk of bias and discrimination, decontextualization, lower academic integrity, or plagiarism. Most of these come to bear in AI-chatbot interaction for learning including for researching information during assessment.

Despite limitations, our study is among the first to investigate both students' assessment performance and real-world learning outcomes in relation to chatbot use in higher education. By explicitly differentiating between learning quality (outcomes) and speed (time to completion), it contributes to a to a more nuanced understanding of chatbot effects than most short-term studies (Wang and Fan 2025; Heung and Chiu 2025). This distinction directly speaks to current discussions on how AI affects the validity of assessment, particularly the concern that AI support may inflate apparent performance without corresponding gains in conceptual understanding (Zhao et al. 2024).

The findings suggest that while chatbots may increase efficiency, their contribution to deeper learning remains uncertain. These results are in line with previous

research (Elnaffar et al. 2025, Laun and Wolff 2025), though highlight the need to distinguish between the speed and quality of learning when evaluating the role of AI tools in higher education. Faster task completion does not necessarily indicate improved understanding or higher achievement. This adds empirical weight to emerging debates around whether AI-supported tasks should be treated differently in formative versus summative assessment, where speed may be beneficial in the former but problematic in the latter (Luo et al. 2025).

Pedagogically, this implies that students should be supported in developing metacognitive and critical reasoning strategies to effectively monitor and evaluate chatbot outputs (Alsaiari et al. 2025; Singh et al. 2025; Tankelevitch et al. 2024; Winne and Azevedo 2022). From an institutional perspective, curricula may need to integrate chatbot literacy and critical AI use (Jin et al. 2025; Rojas-Contreras et al. 2025; Holmes and Miao 2023; Chen 2025), focusing not only on technical proficiency but also on epistemic vigilance and quality assurance (Qu et al. 2025; Fulsher et al. 2025; McPhee and Jerowsky 2025). Our findings reinforce these ongoing discussions by indicating that students cannot rely solely on chatbot outputs for academically credible work, supporting calls for explicit training in AI-supported reasoning and verification.

Conceptually, these findings underline the necessity of models that account for both behavioral and cognitive dimensions of chatbot use over time (de Mooij et al. 2025; Taub et al. 2022), rather than relying on usage frequency as a sole predictor of learning outcomes (Khosravi et al. 2025). This aligns with current debates about moving from tool-centric to pedagogy-first frameworks that emphasize how AI affects learning processes rather than simply documenting tool usage (Luo et al. 2025). A more detailed conceptualization of similarities and differences between COR research with and without AI needs to be undertaken and tested using process data. Further longitudinal, domain-sensitive research is needed to examine the dual roles of AI-tools in shaping both the effectiveness and the pace of learning in realistic educational settings (Vieriu and Petrea 2025). Such work is increasingly seen as essential to guiding evidence-informed policy on AI-supported assessment (Zhao et al. 2024).

Supplementary Information The online version of this article (https://doi.org/10.1007/s42010-026-00242-2) contains supplementary material, which is available to authorized users.

Acknowledgements This research was funded as part of the research unit (FOR) CORE 5404 under grant number 462702138. We sincerely thank the editors and anonymous reviewers for their invaluable and constructive feedback on earlier versions of this paper.

Funding This work is part of the research unit (FOR) 5404 CORE (Critical Online Reasoning in Higher Education) (2023–2027) funded by the German Research Foundation (DFG).

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest D. Molerov, D. Federiakin, O. Zlatkin-Troitschanskaia, K. Shenavai, L. Trierweiler and M.-T. Nagel declare that they have no competing interests.

References

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, *21*(1), 10.

Abdaljaleel, M., Barakat, M., Alsanafi, M., Salim, N. A., Abazid, H., Malaeb, D., Mohammed, A. H., Hassan, B. A. R., Wayyes, A. M., Farhan, S. S., Khatib, S. E., Rahal, M., Sahban, A., Abdelaziz, D. H., Mansour, N. O., AlZayer, R., Khalil, R., Fekih-Romdhane, F., Hallit, R., Hallit, S., & Sallam, M. (2024). A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. *Scientific Reports*, *14*(1), 1983. https://doi.org/10.1038/s41598-024-52549-8.

Adelegan, J. (2024). *The impact of ChatGPT on students' performance*. Lappeenranta-Lahti University of Technology LUT. Bachelor's thesis

Akbar, M. S. (2025). *Beyond detection: designing AI-resilient assessments with automated feedback tool to foster critical thinking*. arXiv:2503.23622. https://arxiv.org/abs/2503.23622

Al-Zahrani, A. M., & Alasmari, T. M. (2024). Exploring the impact of artificial intelligence on higher education: the dynamics of ethical, social, and educational implications. *Humanities and Social Science Communications*, *11*, 912. https://doi.org/10.1057/s41599-024-03432-4.

Alexander, P. A. (2004). A model of domain learning: reinterpreting expertise as a multidimensional, multistage process. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition* (pp. 287–312). Routledge.

Almogren, A. S., Al-Rahmi, W. M., & Dahri, N. A. (2024). Exploring factors influencing the acceptance of ChatGPT in higher education: A smart education perspective. *Heliyon*, *10*(11), e31887. https://doi.org/10.1016/j.heliyon.2024.e31887.

Almulla, M. A. (2024). Investigating influencing factors of learning satisfaction in AI ChatGPT for research: university students perspective. *Heliyon*, *10*(11), e32220. https://doi.org/10.1016/j.heliyon.2024.e32220.

Alsaiari, O., Baghaei, N., Lodge, J. M., Noroozi, O., Gasevic, D., Boden, M., & Khosravi, H. (2025). *Directive, metacognitive or a blend of both? A comparison of AI-generated feedback types on student engagement, confidence, and outcomes*. arXiv:2510.19685. https://arxiv.org/abs/2510.19685

Alshammari, S. H., & Alshammari, M. H. (2024). Factors affecting the adoption and use of ChatGPT in higher education. *International Journal of Information and Communication Technology Education (IJICTE)*, *20*(1), 1–16. https://doi.org/10.4018/IJICTE.339557.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., & Mariman, R. (2024). *Generative AI can harm learning*. https://doi.org/10.2139/ssrn.4895486. SSRN preprint

Behm, T. (2018). *SWE-IV-16. Skala zur Erfassung der Informationsverhaltensbezogenen Selbstwirksamkeitserwartung*. Trier: Leibniz-Institut für Psychologie (ZPID), Open Test Archive. Verfahrensdokumentation, Fragebogen deutsche und englische Version (SES-IB-16)

Bentler, P. M. (1968). Alpha-maximized factor analysis (alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika*, *33*(3), 335–345. https://doi.org/10.1007/BF02289328.

Biesta, G. (2007). Why "what works" won't work: evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x.

Bozkurt, A., Xiao, J., Farrow, R., Bai, J. Y., Nerantzi, C., Moore, S., Dron, J., Stracke, C. M., Singh, L., Crompton, H., Koutropoulos, A., Terentev, E., Pazurek, A., Nichols, M., Sidorkin, A., Costello, E., Watson, S., Mulligan, D., Honeychurch, S., Hodges, C. B., & Asino, T. I. (2024). The manifesto for teaching and learning in a time of generative AI: a critical collective stance to better navigate the future. *Open Praxis*, *16*(4), 487–513.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using Internet. *Computers and Education*, *53*(4), 1207–1217. https://doi.org/10.1016/j.compedu.2009.06.004.

Braun, H., Shavelson, R., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical thinking: conceptualization, design, and implementation. *Frontiers in Education*, *5*, 156. https://doi.org/10.3389/feduc.2020.00156.

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). *The GenAI divide: state of AI in business 2025*. MIT Press. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Chen, B. (2025). *Beyond tools: generative AI as epistemic infrastructure in education*. arXiv:2504.06928. https://arxiv.org/abs/2504.06928

Chiu, Y. L., Liang, J. C., & Tsai, C. C. (2013). Internet-specific epistemic beliefs and self-regulated learning in online academic information searching. *Metacognition and Learning*, *8*(3), 235–260. https://doi.org/10.1007/s11409-013-9103-x.

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, *24*(3), 467–480. https://doi.org/10.1177/001316446402400303.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318.

Drosos, I., Sarkar, A., Xu, X., & Toronto, N. (2025). *"It makes you think": provocations help restore critical thinking to AI-assisted knowledge work*. arXiv:2501.17247. https://arxiv.org/abs/2501.17247

Duenas, T., & Ruiz, D. (2024). *The risks of human overreliance on large language models for critical thinking*. https://doi.org/10.13140/RG.2.2.26002.06082. Research Gate preprint

Elnaffar, S., Rashidi, F., & Abualkishik, A. Z. (2025). *Teaching with AI: a systematic review of chatbots, generative tools, and tutoring systems in programming education*. arXiv:2510.03884. https://arxiv.org/abs/2510.03884

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations. Research report. https://files.eric.ed.gov/fulltext/ED315423.pdf.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846.

Faust, A., Dröge, M., & Odebrecht, C. (2023). Assessment of AI literacy—development and testing of a customizable set of items. In M. Klein, D. Krupka & C. Winter (Eds.), *INFORMATIK 2023—Designing Futures* (pp. 34–346). Gesellschaft für Informatik. https://doi.org/10.18420/inf2023_31.

Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education, 9,* 1366434. https://doi.org/10.3389/feduc.2024.1366434.

Fulsher, A., Pagkratidou, M., & Kendeou, P. (2025). GenAI and misinformation in education: a systematic scoping review of opportunities and challenges. *AI & Society*. https://doi.org/10.1007/s00146-025-02536-y.

Gerlich, M. (2025). AI tools in society: impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6. https://doi.org/10.3390/soc15010006.

Gonsalves, C. (2024). Generative AI's impact on critical thinking: revisiting Bloom's taxonomy. *Journal of Marketing Education*. https://doi.org/10.1177/02734753241305980.

Grassini, S. (2023). Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, *13*(7), 692.

Grassini, S., Aasen, M. L., & Møgelvang, A. (2024). Understanding university students' acceptance of ChatGPT: insights from the UTAUT2 model. *Applied Artificial Intelligence*, *38*(1), e2371168. https://doi.org/10.1080/08839514.2024.2371168.

Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, *100*(12), 4876–4883.

Hair, J. F., Sarstedt, M., Ringle, C. M., Sharma, P. N., & Liengaard, B. D. (2024). Going beyond the untold facts in PLS–SEM and moving forward. *European Journal of Marketing*, *58*(13), 81–106.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Belonax, T., Troy, K.K., Amodci, D., Kaplan, J., Clark, J., & Ganguli, D. (2025). *Which economic tasks are performed with AI? Evidence from millions of Claude conversations*. https://assets.anthropic.com/m/2e23255f1e84ca97/original/Economic_Tasks_AI_Paper.pdf Anthropic preprint.

Hartig, J., Drake, P., Illmann, J., Köhler, C., & Kohmer, A. (in press). Rating procedures to evaluate generic critical online reasoning in an open Internet environment. *Zeitschrift für Erziehungswissenschaft*.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*, 18617. https://doi.org/10.1038/s41598-023-45644-9.

Heung, Y. M. E., & Chiu, T. K. F. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence, 8*, 100361. https://doi.org/10.1016/j.caeai.2025.100361.

Ho, D., Imai, K., King, G., Stuart, E., Whitworth, A., & Greifer, N. (2023). *MatchIt: nonparametric preprocessing for parametric causal inference. CRAN*. https://cran.r-project.org/web/packages/MatchIt/index.html

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO.

Jahn, D., & Kenner, A. (2017). Critical thinking in higher education: how to foster it using digital media. In *The digital turn in higher education: International perspectives on learning and teaching in a changing world* (pp. 81–109). Springer.

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: the generative AI literacy assessment test. *Computers and Education: Artificial Intelligence*, *9*, 100436. https://doi.org/10.1016/j.caeai.2025.100436.

Kaiss, W., Mansouri, K., & Poirier, F. (2024). Effectiveness of chatbots in improving learning experience in an online course. *Interactive Learning Environments*, *33*(3), 2479–2497. https://doi.org/10.1080/10494820.2024.2412059.

Kalai, A.T., Nachum, O., Vempala, S.S., & Zhang, E. (2025). *Why language models hallucinate*. arXiv:2509.04664. https://arxiv.org/abs/2509.04664

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Kelly, R. (2024). *Survey: 86% of students already use AI in their studies. Campus Technology*. https://campustechnology.com/articles/2024/08/28/survey-86-of-students-already-use-ai-in-their-studies.aspx#:~:text=On%20average%2C%20surveyed%20students%20use,Search%20for%20information%20(69%25)%3B

Khosravi, H., Shibani, A., Jovanovic, J., Pardos, Z., & Yan, L. (2025). Generative AI and learning analytics: pushing boundaries, preserving principles. *Journal of Learning Analytics*, *12*(1), 1–11. https://doi.org/10.18608/jla.2025.8961.

Kim, Y., & Steiner, P. (2016). Quasi-experimental designs for causal inference. *Educational Psychologist*, *51*(3-4), 395–405. https://doi.org/10.1080/00461520.2016.1207177.

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J.M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225.

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.H., Beresnitzky, A.V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an ai assistant for essay writing task*. arXiv:2506.08872. https://arxiv.org/abs/2506.08872

Kozyreva, A., Wineburg, S., Lewandowsky, S., & Hertwig, R. (2023). Critical ignoring as a core competence for digital citizens. *Current Directions in Psychological Science*, *32*(1), 81–88.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*(1), 56. https://doi.org/10.1186/s41239-023-00426-1.

Laun, M., & Wolff, F. (2025). Chatbots in education: hype or help? A meta-analysis. *Learning and Individual Differences*, *119*, 102646.

Leighton, J., Cui, Y., & Cutumisu, M. (2021). Key information processes for thinking critically in data-rich environments. *Frontiers in Education*, *6*, 561847. https://doi.org/10.3389/feduc.2021.561847.

Liang, J., Wang, L., Luo, J., Yan, Y., & Fan, C. (2023). The relationship between student interaction with generative artificial intelligence and learning achievement: serial mediating roles of self-efficacy

and cognitive engagement. *Frontiers in Psychology*, *14*, 1285392. https://doi.org/10.3389/fpsyg.2023.1285392.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation—A discussion and demonstration of basic features. *PLoS ONE*, *14*(7), e219854. https://doi.org/10.1371/journal.pone.0219854.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171.

Luo, J., Zheng, C., Yin, J., & Teo, H.H. (2025). Design and assessment of AI-based learning tools in higher education: a systematic review. *International Journal of Educational Technology in Higher Education*, *22*, 42. https://doi.org/10.1186/s41239-025-00540-2.

McDonald, R.P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*, 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x.

McPhee, S.W., & Jerowsky, M. (2025). Beyond technical skills: a pedagogical perspective on fostering critical engagement with generative AI in university classrooms. *Frontiers in Education*, *10*, 1593278. https://doi.org/10.3389/feduc.2025.1593278.

Mike, N., Karsai, K., Orbán, G., Bubelényi, A., Nagy, C.N., & Polyák, G. (2024). The impact of ChatGPT on student performance in higher education. In *Proceedings EGOV-CeDEM-ePart conference*. Ghent University and KU Leuven, 1–5 Sep. CEUR-Workshop Proceedings, (Vol. 3737, pp. 1–15). https://ceur-ws.org/Vol-3737/paper28.pdf.

Minh, A.N. (2024). Leveraging ChatGPT for enhancing English writing skills and critical thinking in university freshmen. *Journal of Knowledge Learning and Science Technology*, *3*(2), 51–62. https://doi.org/10.60087/jklst.vol3.n2.p62.

Mislevy, R.J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mohamed, A., Assi, M., & Guizani, M. (2025). *The impact of LLM-assistants on software developer productivity: a systematic literature review*. arXiv:2507.03156. https://arxiv.org/abs/2507.03156

Molerov, D., Zlatkin-Troitschanskaia, O., Nagel, M. T., Brückner, S., Schmidt, S., & Shavelson, R. J. (2020). Assessing university students' critical online reasoning ability: A conceptual and assessment framework with preliminary evidence. *Frontiers in Education, 5,* 577843. https://doi.org/10.3389/feduc.2020.577843.

de Mooij, S., Lämsä, J., Lim, L., Aksela, O., Athavale, S., Bistolfi, I., Jin, F., Li, T., Azevedo, R., Bannert, M., Gasevic, D., Järvelä, S., & Molenaar, I. (2025). A systematic review of self-regulated learning through integration of multimodal data and artificial intelligence. *Educational Psychology Review*, *37*, 54. https://doi.org/10.1007/s10648-025-10028-0.

Morell-Mengual, V., Fernández-García, O., Berenguer, C., Ortega-Barón, J., Gil-Llario, M.D., & Estruch-García, V. (2025). Characteristics, motivations and attitudes of students using ChatGPT and other language model-based chatbots in higher education. *Education and Information Technologies*, *3*, 22257–22274. https://doi.org/10.1007/s10639-025-13650-1.

Moustaghfir, S., & Brigui, H. (2024). Navigating critical thinking in the digital era: an informative exploration. *International Journal of Linguistics, Literature and Translation*, *7*(1), 137–143. https://doi.org/10.32996/ijllt.2024.7.1.11x.

Nagel, M.T., Zlatkin-Troitschanskaia, O., & Fischer, J. (2022). Validation of newly developed tasks for the assessment of generic critical online reasoning (COR) of university students and graduates. *Frontiers in Education*, *7*, 914857. https://doi.org/10.3389/feduc.2022.914857.

Nagel, M.T., Zlatkin-Troitschanskaia, O., Martin de los Santos Kleinz, L., Braunheim, D., Fischer, J., Maur, A., Shenavai, K., & Kohmer, A. (2024). Critical online reasoning among young professionals: overview of demands and skills in the domains of law, medicine, and teaching. In O. Zlatkin-Troitschanskaia, M.T. Nagel, V. Klose & A. Mehler (Eds.), *Students', graduates' and young professionals' critical use of Online information*. Springer. https://doi.org/10.1007/978-3-031-69510-0_1.

Open, A. (2023). *GPT-4 technical report. (arXiv:2303.08774). (version 1)*. https://arxiv.org/abs/2303.08774

Oser, F.K., & Biedermann, H. (2020). A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis. In *Frontiers and advances in positive learning in the age of InformaTiOn (PLATO)* (pp. 89–106). Springer.

Pan, Z., Xie, Z., Liu, T., & Xia, T. (2024). Exploring the key factors influencing college students' willingness to use AI coding assistant tools: an expanded technology acceptance model. *Systems*, *12*(5), 176. https://doi.org/10.3390/systems12050176.

de la Puente, M., Torres, J., Troncoso, A. L. B., Meza, Y. Y. H., & Carrascal, J. X. M. (2024). Investigating the use of ChatGPT as a tool for enhancing critical thinking and argumentation skills in international relations debates among undergraduate students. *Smart Learning Environments*, *11*(1), 55. https://doi.org/10.1186/s40561-024-00347-0.

Putra, I. M., & Abdunnafi, F. I. (2024). *Understanding the unexpected: the adverse influence of large language models on critical thinking and professional skepticism in accounting education*. https://ssrn.com/abstract=4914889 SSRN preprint.

Qu, X., Sherwood, J., Liu, P., & Aleisa, N. (2025). Generative AI tools in higher education: a meta-analysis of cognitive impact. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. CHI EA '25, April. (Vol. 302, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3706599.3719841.

Rejeleene, R., Xu, X., & Talburt, J. (2024). *Towards trustable language models: Investigating information quality of large language models*. arXiv:2401.13086. https://arxiv.org/abs/2401.13086

Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., Pascual, R. J., Bradley, N., Thomas, D., & Leneghan, F. (2024). ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. *International Journal for Educational Integrity*, *20*, 18. https://doi.org/10.1007/s40979-024-00161-8.

Rodgers, W. M., Markland, D., Selzler, A. M., Murray, T. C., & Wilson, P. M. (2014). Distinguishing perceived competence and self-efficacy: an example from exercise. *Research Quarterly for Exercise and Sport*, *85*(4), 527–539. https://doi.org/10.1080/02701367.2014.961050.

Rojas-Contreras, M., Quintana, L. C., Villamizar, O. A., Bejarano, J. L. M., & Carrillo, L. V. (2025). *Frameworks for artificial intelligence literacy in higher education*. 23rd LACCEI International Multi-Conference for Engineering, Education, and Technology: "Engineering, Artificial Intelligence, and Sustainable Technologies in service of society"., Mexico City, July 16–18, 2025. https://doi.org/10.18687/LACCEI2025.1.1.1620.

Romero-Rodríguez, J. M., Ramírez-Montoya, M. S., Buenestado-Fernández, M., & Lara-Lara, F. (2023). Use of ChatGPT at university as a tool for complex thinking: students' perceived usefulness. *Journal of New Approaches in Educational Research*, *12*(2), 323–339. https://doi.org/10.7821/naer.2023.7.1458.

Royer, C. (2024). Outsourcing humanity? ChatGPT, critical thinking, and the crisis in higher education. *Studies in Philosophy and Education*, *43*, 479–497. https://doi.org/10.1007/s11217-024-09946-3.

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., & Chadha, A. (2024). *A comprehensive survey of hallucination in large language, image, video and audio foundation models*. arXiv:2405.09589. https://arxiv.org/abs/2405.09589

Salas-Martínez, Á., & Ramírez-Martinell, A. (2025). Learning analytics in higher education: a decade in systematic literature review. *Programming and Computer Software*, *50*, 690–700. https://doi.org/10.1134/S0361768824700701.

Schei, O. M., Møgelvang, A., & Ludvigsen, K. (2024). Perceptions and use of AI chatbots among students in higher education: a scoping review of empirical studies. *Education Sciences*, *14*(8), 922. https://doi.org/10.3390/educsci14080922.

Schemer, C., Matthes, J., & Wirth, W. (2008). Toward improving validity and reliability of information processing measures in surveys. *Communication Methods and Measures*, *2*(2), 1–33.

Scherr, S., Cao, B., Jiang, L. C., & Kobayashi, T. (2025). Explaining the use of AI chatbots as context alignment: motivations behind the use of AI chatbots across contexts and culture. *Computers in Human Behavior*, *172*, 108738.

Schmidt, S., Zlatkin-Troitschanskaia, O., Roeper, J., Klose, V., Weber, M., Bültmann, A.-K., & Brückner, S. (2020). Undergraduate students' critical online reasoning—process mining analysis. *Frontiers in Psychology*, *11*, 576273. https://doi.org/10.3389/fpsyg.2020.576273.

Shaengchart, Y., Bhumpenpein, N., Kongnakorn, K., Khwannu, P., Tiwtakul, A., & Detmee, A. (2024). Factors influencing the acceptance of ChatGPT usage among higher education students in Bangkok, Thailand. *Advance Knowledge for Executives*, *2*(4), 1–14. https://ssrn.com/abstract=4592118.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *International Journal of Testing*, *19*(4), 337–362.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257.

Singh, A., Guan, Z., & Rieh, S. Y. (2025). Enhancing critical thinking in generative AI search with metacognitive prompts. *Proceedings of the Association for Information Science and Technology*, *62*(1), 672–684.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC. https://doi.org/10.1201/9780203489437.

Sparks, J. R., Ober, T. M., Tenison, C., Arslan, B., Roll, I., Deane, P., Zapata-Rivera, D., Gooch, R., & O'Reilly, T. (2024). *Opportunities and challenges for assessing digital and AI literacies [report]*. Educational Testing Service. https://www.ets.org/pdfs/rd/ets-digital-literacy-ai-full-report.pdf

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386. https://doi.org/10.1016/j.chb.2024.108386.

Stöhr, C., Ou, A. W., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels, and fields of study. *Computers and Education: Artificial Intelligence*, *7*, 100259. https://doi.org/10.1016/j.caeai.2024.100259.

Strzelecki, A. (2024). Students' acceptance of ChatGPT in higher education: an extended unified theory of acceptance and use of technology. *Innovative Higher Education*, *49*(2), 223–245. https://doi.org/10.1007/s10755-023-09686-1.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), 84–90. https://doi.org/10.1016/j.jclinepi.2013.01.013.

Swargiary, K. (2024). *The impact of ChatGPT on student learning outcomes: a comparative study of cognitive engagement, procrastination, and academic performance*. https://doi.org/10.2139/ssrn.4914743. SSRN preprint

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Article 680. (pp. 1–24). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642902.

Taub, M., Banzon, A. M., Zhang, T., & Chen, Z. (2022). Tracking changes in students' online self-regulated learning behaviors and achievement goals using trace clustering and process mining. *Frontiers in Psychology*, *13*, 813514. https://doi.org/10.3389/fpsyg.2022.813514.

Teo, T. (Ed.). (2011). *Technology acceptance in education: research and issues*. Sense Publishers. https://doi.org/10.1007/978-94-6091-487-4.

Tsang, R., & Wood, S. Y. (2025). *Help or hype? Exploring LLM-based chatbots in self-regulated learning*. 2025 ASEE Annual Conference & Exposition, Montreal, June 22–25. American Society for Engineering Education. https://doi.org/10.18260/1-2--56687.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2016). Unified theory of acceptance and use of technology: a synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328–376. https://ssrn.com/abstract=2800121.

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, *15*(3), 343. https://doi.org/10.3390/educsci15030343.

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 1–21. https://doi.org/10.1057/s41599-025-04787-y.

Wang, K. D., Salehi, S., & Wieman, C. (2023). Applying log data analytics to measure problem solving in simulation-based learning environments. In V. Kovanovic, R. Azevedo, D. C. Gibson & D. Ifenthaler (Eds.), *Unobtrusive observations of learning in digital environments. Advances in analytics for learning and teaching*. Springer. https://doi.org/10.1007/978-3-031-30992-2_3.

Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education*, *8*, 1330486. https://doi.org/10.3389/feduc.2023.1330486.

Wineburg, S., Breakstone, J., McGrew, S., Smith, M. D., & Ortega, T. (2022). Lateral reading on the open Internet: a district-wide field study in high school government classes. *Journal of Educational Psychology*, *114*(5), 893–909. https://doi.org/10.1037/edu0000740.

Winne, P. H., & Azevedo, R. (2022). Metacognition and self-regulated learning. In R. K. Sawyer (Ed.), *The cambridge handbook of the learning sciences*. Cambridge handbooks in psychology. (pp. 93–113). Cambridge University Press.

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, *55*(1), 10–33.

Youssef, E., Medhat, M., Abdellatif, S., & Al Malek, M. (2024). Examining the effect of ChatGPT usage on students' academic learning and achievement: A survey-based study in Ajman, UAE. *Computers and Education: Artificial Intelligence*, *7*, 100316.

Yu, C., Yan, J., & Cai, N. (2024). ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention. *Frontiers in Education*, *9*, 1354929. https://doi.org/10.3389/feduc.2024.1354929.

Yudytska, J., & Androutsopoulos, J. The use of language technologies in forced migration: an explorative study of Ukrainian women in Austria. In M. Mendes de Oliveira & L. Conti (Eds.), *Explorations in digital interculturality: language, culture, and postdigital practices*. transcript. in press.

Zakir, S., Hoque, M., Susanto, P., Nisaa, V., Alam, M., Khatimah, H., & Mulyani, E. (2025). Digital literacy and academic performance: the mediating roles of digital informal learning, self-efficacy, and students' digital competence. *Frontiers in Education*, *10*, 1590274. https://doi.org/10.3389/feduc.2025.1590274.

Zhao, J., Chapman, E., & Ghassemi Pour Sabet, P. (2024). Generative AI and educational assessments: a systematic review. *Education Research & Perspectives*. https://doi.org/10.70953/ERPv51.2412006.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*(8032), 61–68. https://doi.org/10.1038/s41586-024-07930-y.

Zirar, A. (2023). Exploring the impact of language models, such as ChatGPT, on student learning and assessment. *Review of Education*, *11*(3), e3433. https://doi.org/10.1002/rev3.3433.

Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., & Lautenbach, C. (2020). *Student learning in German higher education*. Springer.

Zlatkin-Troitschanskaia, O., Hartig, J., & König, J. (2025). Special issue: Critical online reasoning in higher education: Conceptual and methodological advances and perspectives in internet-based learning (Guest editors). *Zeitschrift für Erziehungswissenschaft*.