



Enhancing School Students' Self-Regulated Learning through Generative AI Support: A Randomized Controlled Trial

Tim Fütterer¹ · Lisa Bardach^{2,3} · Jochen Kuhn⁴ · Stefan Daniel Keller⁵ · Peter Gerjets⁶

Received: 30 June 2025 / Accepted: 12 March 2026
© The Author(s) 2026

Abstract

Self-regulated learning (SRL) is a critical competency for academic success and lifelong learning, yet providing personalized SRL support in real-time classroom settings remains challenging. In this randomized controlled trial, we investigated whether generative-AI-(GenAI)-based interventions can effectively support SRL in secondary school students. Drawing on an integrative model of self-regulation in education, the study contrasted two types of GenAI-supported interventions—targeting either motivational (utility value) or strategic (cognitive-learning-strategy-based) aspects of SRL—with a control condition using standard ChatGPT. A total of 371 students (Grades 7–9) were randomly assigned to one of the three conditions and participated in six 45-min sessions during regular physics or English lessons. Outcome measures included self-reported utility value, strategy use (self-reported and tested performance), interest, effort, and tested domain-specific knowledge, assessed at pre- and posttest. The results showed that students in the utility value condition reported more favorable development of perceived utility value than those in the cognitive strategy condition. However, no statistically significant advantages of either intervention over the control condition were found for effort, domain-specific knowledge, or elaboration-based strategy use. Exploratory analyses indicated that students who engaged more meaningfully with the GenAI tended to have more sustained interest than those in the control group. No subject-specific (physics or English) differences in intervention effects were observed. Findings suggest GenAI-based interventions may help preserve motivational aspects of SRL under certain conditions, but further development is needed to effectively support cognitive strategies and improve learning outcomes in secondary school settings.

Extended author information available on the last page of the article

Published online: 14 April 2026

Springer

Highlights

- GenAI-based tools were tested for promoting SRL in a school-based RCT design.
- Utility value prompts helped preserve students' utility value in the learning content.
- No clear condition differences emerged for other learning outcomes.
- Exploratory analyses revealed that meaningful GPT interactions buffered motivational declines.
- The study provides ideas for scalable and classroom-integrated SRL support using GenAI.

Keywords Self-regulated learning · Generative AI · Utility value · Cognitive strategies · Randomized controlled trial · Secondary education

Self-regulated learning (SRL) skills are crucial for effective learning (Dent & Koenka, 2016; Geng et al., 2024), as SRL enables students to take control of their learning processes by setting goals, monitoring progress, and adjusting strategies accordingly (Zimmerman, 2002). Given its importance for academic success and lifelong learning, fostering SRL is a major goal in educational research and practice (Deutsche Akademie Der Naturforscher Leopoldina, 2024; Foster & Piacentini, 2023; Schunk & Zimmerman, 1998). However, due to its complexity, SRL requires support mechanisms to help students engage effectively in self-regulatory processes (Boekaerts & Corno, 2005; Dignath & Büttner, 2008; Zimmerman, 2002). SRL has been described by numerous models, which can be used to identify effective intervention points (Tinajero et al., 2024). Early models, such as those by Zimmerman (2002), Pintrich (2005), Boekaerts (1996), Winne and Hadwin (1998), and Efklides (2011), conceptualized SRL as involving multiple interrelated cognitive, metacognitive, and motivational components, although they differed in their emphasis on dynamic sequencing, affective processes, and levels of processing.

A promising integrative framework is offered by the model by Wortha et al. (2026, see this special issue), which situates SRL within the broader context of self-regulation in education. This model distinguishes four key components that interact to shape learning outcomes: *personal dispositions* (e.g., personality traits such as conscientiousness) that describe learners' constraints when entering a learning episode, *driving forces* (e.g., current motivation) that energize learning behaviors, *limited resources* (e.g., cognitive capacity, executive functions) that determine the efficiency of SRL processes, and *learning activities* (e.g., cognitive learning strategies) that represent direct engagement with learning tasks. These components are supposed to interact in complex, partially compensatory ways. For instance, students with strong driving forces (e.g., high levels of motivation) may be able to compensate for limited cognitive resources due to effort investment. By contrast, students who have elaborated strategy knowledge may nevertheless struggle to apply sophisticated strategies if their executive functioning is insufficient for their deployment.

Although several theoretical advancements have occurred in recent years (Tinajero et al., 2024), implementing effective SRL support in real classrooms remains a

considerable challenge. Teachers often lack the capacity and competencies to provide individualized SRL support for all students in real time, as such support would require ongoing assessments of students' motivational states, cognitive resources, and strategy use, as well as the use of these assessments for the selection of appropriate learning tasks and support measures (Azevedo & Aleven, 2013; Fütterer et al., 2024, 2026; Greene & Azevedo, 2010; Praetorius et al., 2025; see also Plass & Pawar, 2020 for adaptive teaching). This area is where generative artificial intelligence (GenAI) may have great potential. On the one hand, GenAI-driven tools may offer scalable and personalized support for reducing teachers' workloads (Fütterer et al., 2023; Kasneci et al., 2023; Steinert et al., 2024), for instance, when it comes to helping students deploy effective cognitive learning strategies or fostering their motivation (e.g., utility values). On the other hand, existing research on the use of GenAI tools in education—particularly in K–12 contexts—often suffers from methodological limitations, theoretical inconsistencies, and a lack of experimental designs (X. Han et al. 2025a, b). Thus, more reliable evidence demonstrating the potential of GenAI tools for SRL support in learning contexts is needed.

Against this backdrop, we conducted the present study to implement a theory-informed randomized controlled trial (RCT) in real classrooms, with clearly defined treatments and outcome measures. Thus, the present study was designed to contribute not only to theory and research but also to practice in the field of SRL. Theoretically, we drew on the integrative framework by Wortha et al. (2026) to examine GenAI-supported interventions targeting either motivational (utility value) or strategic (cognitive-learning-strategy-based) components of SRL. Empirically, we investigated the effects of these CustomGPT-based interventions on motivation, strategy use, effort, and learning outcomes. Practically, we explored the potential of GenAI tools to provide scalable and personalized SRL support in educational settings.

Theoretical Background

A Comprehensive Understanding of Self-Regulation

Traditional models of SRL have considerably shaped the understanding of how learners (should) plan, monitor, and regulate their learning processes (e.g., Boekaerts, 2007; Pintrich, 2005; Winne & Hadwin, 1998; Zimmerman, 2002). Despite differences in emphases, most models converge on several core assumptions. First, SRL is a goal-directed activity that requires learners to manage internal and external resources. Second, it is context-sensitive, making it responsive to situational affordances and constraints. Third, effective SRL requires learners to dynamically coordinate various strategies, including attention regulation, cognitive engagement, motivational control, and emotional regulation (see Geng et al., 2024).

Although existing SRL models remain influential, recent theoretical developments have moved toward more integrative, multidimensional, and multidisciplinary views of self-regulation that increasingly account for “hot” self-regulation, implicit processes, and situational influences (Tinajero et al., 2024). Specifically, Wortha (2021) and Wortha et al. (2026) proposed a framework that positions SRL within

a general account of self-regulation in education. This integrative framework synthesizes perspectives from educational psychology, motivation science, personality psychology, and cognitive neuroscience by categorizing existing research programs into four broad traditions, each of which is characterized by a core explanatory family of constructs, namely, (a) *personal dispositions*, (b) *driving forces*, (c) *limited resources*, and (d) *learning activities* (see Fig. 1). These traditions differ in several key respects, including their assumptions about the stability and malleability of self-regulation processes, their preferred methodological approaches, and their (temporal) level of analysis.

First, research grounded in *personal-disposition* perspectives suggests that (relatively) stable learner characteristics are important antecedents of successful learning. Typical examples in educational contexts include the Big Five personality traits (e.g., John et al., 2008), dispositional interests (e.g., Hidi & Renninger, 2006; Renninger & Hidi, 2022), and academic self-concepts (e.g., Marsh 1990a, b; Möller 2024; Shavelson et al. 1976). These traits represent broad, long-term influences that help explain stable patterns in students' SRL across tasks and time, shaping how learners typically approach academic demands. Personal dispositions shape how people typically engage with academic demands by influencing learners' likelihood of adopting certain behaviors, thought patterns, and emotional responses. Understanding how these more distal personal dispositions interact with more proximal self-regulatory processes is crucial for developing a comprehensive framework of self-regulation in educational settings.

Second, explanations based on *driving forces* emphasize motivational and emotional processes that energize engagement with learning tasks. Although SRL is often framed in cognitive and metacognitive terms, effective regulation extends beyond

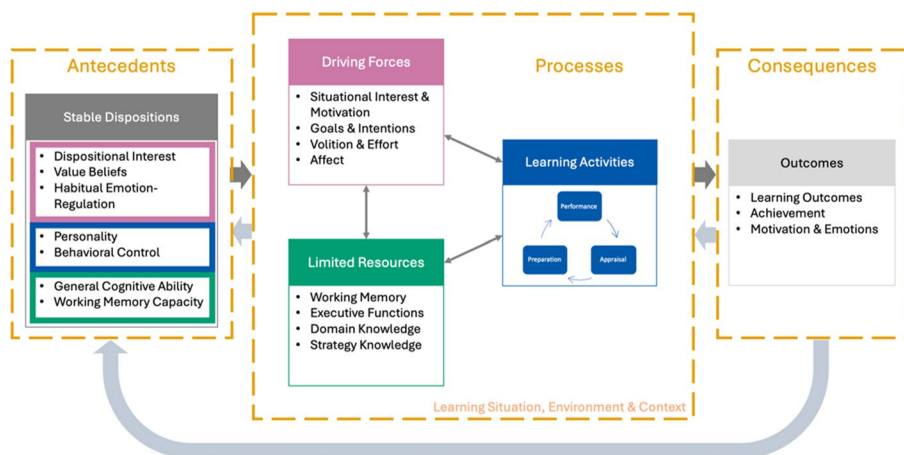


Fig. 1 An integrative framework for self-regulation in education. *Note.* Arrows indicate hypothesized relationships between framework components. Darker arrows represent temporal and causal influences, while lighter arrows denote reciprocal effects. The color-coded mappings between stable dispositions and other parts of the model illustrate primary associations. The lists of concepts within the main framework components are exemplary and not exhaustive. Framework adopted from Wortha et al. (2026; see this special issue)

“cold” cognition and requires processes that activate, energize, and direct learning (Tinajero et al., 2024). For instance, even the availability of the most sophisticated cognitive and metacognitive skills and strategies may fail to produce meaningful outcomes if learners lack the motivation to apply them (see also Zimmerman, 2002). Relevant driving forces in the literature include interconnected constructs, such as expectancy and value beliefs for specific learning goals (e.g., Eccles & Wigfield, 2020, 2023), the types of achievement goals that learners adopt (e.g., Elliot & McGregor, 2001; see also Bardach et al. 2020; Hulleman et al. 2010a, b), achievement emotions (e.g., Pekrun, 2006; Sakaki et al., 2024), as well as situational interest (e.g., Hidi & Renninger, 2006; Knogler et al., 2015; Renninger & Hidi, 2022). Driving forces in learning situations are transient states that differ from more general, stable motivational dispositions. They unfold dynamically within specific learning contexts as learners develop and pursue particular goals. Because SRL is grounded in learners’ active, goal-directed management of their learning (e.g., Schunk & Greene, 2018), a conceptual account of how driving forces such as task interest and affective states shape this engagement is essential (Wortha et al., 2026).

Third, engaging in SRL is inherently resource-demanding, requiring sustained mental effort and strategic attentional focus. Because of its cognitively demanding nature, SRL draws heavily on *limited cognitive resources* (Wortha et al., 2026; Pintrich & Zusho, 2002; Winne, 1995). Executive functions have been described as domain-general, higher level cognitive processes that enable individuals to regulate their thoughts and actions toward goal-directed behaviors by influencing lower level cognitive functions (e.g., perceptual and motor processes; Friedman & Miyake, 2017; Miyake et al., 2000). Yet, despite their importance for learning, limited cognitive resources such as executive functions have received surprisingly little attention in SRL research (Wortha et al., 2026; Wortha, 2021).

Fourth, a key area of research on SRL in education focuses on the actual learning activities students engage in during a learning session, specifically when they prepare, perform, and evaluate their learning tasks and learning strategies (e.g., Zimmerman, 2002). SRL centers on the use, regulation, and adjustment of cognitive and metacognitive learning strategies (e.g., Dignath et al., 2008), which are included in the *learning activities* category in the integrative model of self-regulation in education (Wortha et al., 2026). During the learning phase, cognitive learning activities capture strategic actions learners use to process information, such as rehearsal, elaboration, or critical evaluation. Effective strategies are selected through metacognitive processes and vary by domain and available learning materials (Dunlosky et al., 2013). For instance, generating inferences that link pictorial and verbal information has been shown to support learning and transfer in science multimedia learning (Kühl et al., 2011). In mathematics, problem- and example-based learning benefits from strategies such as problem sequencing and example comparison (Scheiter & Gerjets, 2002). In language learning, rote vocabulary acquisition is enhanced by spacing rather than massed practice (Morehead et al., 2018). Finally, strategies involving retrieval practice and generative activities are often effective for learning (e.g., learning by explaining to others; Lachner et al., 2022). Strategic learning requires students not only to identify appropriate learning strategies for a given situation but also to implement them effectively. Studies on strategy training have shown that this issue is

not trivial. It might not be enough to simply convey strategy knowledge to students (Scheiter et al., 2015); rather, it might also be necessary to help learners implement the strategies they already know (Hoch et al., 2023; Stalbovs et al., 2015).

At its core, the integrative framework by Wortha et al. (2026) conceptualizes self-regulation as a person–situation interaction, wherein stable learner characteristics shape how situation-specific self-regulatory processes unfold and influence educational outcomes (see Fig. 1). This view demands integrating constructs that differ in granularity and temporal scope. Because the framework is grounded in the distinction between different explanatory mechanisms, some dispositional traits may encompass constructs more commonly associated with other traditions. For instance, working memory capacity and well-developed individual interest are stable dispositions, whereas their situational counterparts, working memory load and triggered situational interest, vary with task demands and context.

The integrative framework assumes that the four components have partly compensatory relationships. For instance, highly motivated learners with strong driving forces might compensate for lower cognitive capacities, whereas learners with limited motivation might still succeed if they possess effective cognitive strategies and executive control. Similarly, students with lower cognitive ability may be able to circumvent this limitation to some extent by developing greater specific knowledge, including the use of targeted learning strategies (e.g., Bardach et al., 2025).

By integrating constructs from adjacent research traditions, the framework advances a more comprehensive understanding of self-regulation suited to educational contexts. It may facilitate more accurate diagnostic processes regarding students' achievements and attitudes, thereby informing more targeted domain-specific interventions. Its broad conception of SRL also enables the design of adaptive tools, such as GenAI-based tutors, that respond in real time to learners' motivational, cognitive, and strategic profiles.

Supporting SRL in Accordance with the Comprehensive Model

Several meta-analyses have demonstrated the effectiveness of SRL interventions in promoting students' achievement, strategy use, and motivation across educational contexts and age groups (e.g., Dignath & Büttner, 2008; Jansen et al., 2019). For instance, Dignath et al. (2008) reported a weighted average effect size of $d=0.69$ based on 263 effect sizes, with particularly strong effects on strategy use and motivation among younger learners. SRL interventions have also proven effective in subject-specific domains such as mathematics (Donker et al., 2014), in second language acquisition (Chen, 2022), and in online or blended learning environments (Xu et al., 2023). Overall, these findings emphasize that SRL can be successfully fostered at both the primary and secondary school levels.

This background raises the question of where in the learning process interventions will be most effective. According to the integrative framework of self-regulation in education by Wortha et al. (2026), fostering SRL involves addressing multiple aspects. In this study, we focused on driving forces and learning activities because these dimensions can be directly affected through targeted interventions, such as fostering students' utility value beliefs and thus their situational interest and promoting specific

cognitive strategies for elaborating learning contents (e.g., learning by explaining to others; Lachner et al., 2022). By contrast, personal dispositions (e.g., personality traits) and limited (cognitive) resources are learner characteristics that are less likely to be changed by short-term instructional support. Two effective strategies are (a) emphasizing the value of the learning content (e.g., utility value interventions; Hulleman et al. 2010a, b) and (b) elaborating on it (e.g., by explaining it to others; Donker et al., 2014).

Utility Value Interventions

Among the many motivational interventions in education (Lazowski & Hulleman, 2016), *utility value interventions*—where learners reflect on the personal relevance or usefulness of a topic—have received considerable attention. Utility value interventions have shown particular promise for promoting student interest—a key motivational construct linked to engagement, persistence, and deep learning (Harackiewicz et al. 2016; Hulleman et al. 2010a, b; Renninger and Su 2012). Drawing on the four-phase model of interest development (Hidi & Renninger, 2006; Renninger & Hidi, 2022), interest is understood as progressing from triggered situational interest to maintained situational interest and eventually to a more enduring individual interest (Hulleman & Harackiewicz, 2009). By encouraging learners to connect academic content to their personal goals and lives, utility value interventions can spark initial interest and support their development over time (Harackiewicz et al. 2016; Hecht et al. 2021; Hulleman et al. 2010a, b; Hulleman and Harackiewicz 2009).

Elaboration

In addition to emphasizing the value of the learning content, prompting students to elaborate on it—especially through generative strategies such as explaining—has been shown to be a highly effective cognitive learning strategy (Donker et al., 2014; Fiorella & Mayer, 2015; Lachner et al., 2022). Generative learning strategies require learners to actively construct meaning by integrating new information with their existing knowledge structures (Brod, 2021; Chi, 2000). According to Fiorella's (2023) generative sense-making framework, elaboration-based strategies (e.g., explaining) play a central role in making sense of learning material. The framework posits that learners mentally organize and simulate material (e.g., through visualization and enactment) to enable high-quality verbalizations and generalizations (i.e., explanations), thereby fostering deep and transferable understanding.

Among generative strategies, explaining—particularly explaining to others (see Chi, 2000 for a distinction between self-explanations and explanations for others)—has received growing attention in recent years. The presence of an imagined audience can increase perceived social presence and encourage learners to engage more effortfully and coherently with the material (Fiorella & Mayer, 2015; Holmes, 2007; Lachner et al., 2021). Research has shown that generating instructional explanations to a fictitious student—even in noninteractive settings—can foster conceptual understanding and metacognitive awareness (Lachner et al., 2021). Whereas self-explanation has shown robust effects across domains (Chi, 2000; Hausmann & Vanlehn,

2007), recent findings indicate that explaining to others can yield additional learning gains, especially when learners are guided to generate high-quality explanations (Fiorella, 2022; Zhu et al., 2024).

Potential of GenAI to Foster SRL

Although SRL is essential for academic success, particularly in technology-enhanced environments (M. L. Bernacki et al., 2011; Greene et al., 2010; Greene & Azevedo, 2010), and SRL strategy interventions are effective measures for fostering SRL (e.g., Dignath et al., 2008), the effectiveness of SRL often depends on timely and context-sensitive feedback (Azevedo & Aleven, 2013; Greene & Azevedo, 2010) that teachers cannot always provide for each student (Fütterer et al., 2024). In addition, in classrooms, explicit strategy instruction was found to be rare, although it is associated with performance gains (e.g., compared with implicit strategy instruction; Dignath & Veenman, 2021; Kistner et al., 2010). Teachers might lack knowledge of essential components of SRL and how to foster SRL and thus might be rather reluctant to promote it (Dignath & Büttner, 2018).

AI, especially GenAI, is increasingly seen as a promising way to address this gap by offering scalable, adaptive, and personalized support for learners (Molenaar et al., 2023; OECD, 2023). AI systems can serve both diagnostic and supportive functions. Analytics-based AI approaches (e.g., multimodal learning analytics) can be used to assess learners' SRL processes. For instance, intelligent tutoring systems (ITS; see, Mousavinasab et al., 2021) have been developed to provide structured, theory-based support for SRL (Azevedo et al., 2022). However, ITS development is still relatively resource- and time-intensive, requiring expert-built domain, learner, and pedagogical models to guide instruction and feedback (Kulik & Fletcher, 2016). By contrast, GenAI tools such as large language models (LLMs)—although likely less precise in instructional alignment and associated with challenges such as overreliance, reliability, and ethical concerns (Han et al. 2025a, b)—are more interactive and flexible in adapting to learners' needs through flexible dialogue without requiring extensive pre-modeling and could be used to support SRL by providing real-time, individualized feedback that guides learners' cognitive, motivational, and metacognitive processes (Fütterer et al., 2023; Kasneci et al., 2023; Küchemann et al., 2025). For instance, Brod (2024) argued that recent advances in AI, particularly LLMs, can support the integration of multiple paths to personalized education. Beyond assigning tasks and support levels, AI could also promote SRL by providing personalized feedback and fostering generative learning strategies such as learning by explaining. That is, chatbots can prompt learners to apply cognitive strategies (e.g., summarizing or explaining) or increase motivation by linking content to personal goals and demonstrating its utility for daily life.

Empirical findings on GenAI's impact have been mixed. On the one hand, recent studies have shown the effectiveness of AI in diverse learning contexts (X. Han et al. 2025a, b). For instance, Lim et al. (2023) found that rule-based scaffolding from AI increased students' self-reported metacognitive monitoring while writing, and Cosentino et al. (2025) demonstrated that GPT-4-based feedback during embodied

mathematics learning lowered cognitive load, as assessed using eye-tracking data, and improved visual processing more than teacher-led instruction. In the context of foreign language learning, LLM-based feedback supported both the development of writing competencies and the motivation to engage in more learning (Meyer et al., 2024). It has also been demonstrated that learners who could ask questions—whether interacting with peers or ChatGPT—achieved better post dialogue performance than those who only answered questions posed by an ITS, highlighting the crucial role of question-asking in dialogic learning (Yang et al., 2025). Even though GenAI holds promise for supporting all phases of SRL (Kong & Yang, 2024), a recent qualitative systematic review showed that AI, especially GenAI, were predominantly used to support the performance phase of SRL, with fewer systems targeting forethought and reflection (Lan & Zhou, 2025). Overall, these results suggest that AI, especially GenAI, can foster SRL—provided that systems are designed to preserve learner agency and promote active engagement.

On the other hand, Han et al. (2025a, b) showed that AI-supported chatbots effectively enhanced goal setting, monitoring, and reflection in students with initially low SRL skills, but students with high initial SRL levels demonstrated a decrease in strategy use over time, possibly due to overscaffolding. Similarly, Fan et al. (2025) found that whereas ChatGPT improved essay quality in a university writing task, it led to lower levels of metacognitive processing, indicating potential side effects such as *metacognitive laziness*. To mitigate risks such as overscaffolding, metacognitive laziness, and the potential negative effects of AI, especially GenAI, researchers have proposed hybrid frameworks that integrate human and AI regulation. For instance, Molenaar (2022a, b) introduced the concept of Hybrid Human-AI Regulation (HHAIR), which utilizes dashboards to scaffold SRL and gradually shift control from AI (AI regulation) to the learner (self-regulation), thereby fostering autonomy and long-term learning capacity (see also Järvelä et al., 2025).

Overall, despite the growing popularity of GenAI, such as ChatGPT, in educational practice, scholars have emphasized the need for more robust research designs, such as multitask and cross-context approaches that better reflect the realities of learning across disciplines and settings (Fan et al., 2025). Many current studies have lacked a sound theoretical foundation. Although some have referenced well-established models, such as Zimmerman's model of SRL, cognitive load theory, or self-determination theory, these frameworks have often been applied inconsistently or superficially (Lan & Zhou, 2025). There is also a dearth of RCTs conducted in authentic K–12 school settings, with most studies taking place in higher education contexts (Wu & Yu, 2024). Most importantly, it remains unclear how learners can benefit from AI agents (Fan et al., 2025). Particularly regarding SRL, further research is needed to understand which SRL components can be effectively supported by AI, especially GenAI, and how different types of support (e.g., motivational vs. strategic scaffolding) impact student learning. To address these gaps, the present study implemented a theory-informed, school-based RCT to examine the effects of different types of GenAI-based support on students' SRL.

The Present Study

Despite substantial theoretical progress, providing effective, individualized SRL support in real classrooms remains a major challenge. This study addressed this gap by investigating how GenAI can provide scalable, adaptive support in different subject-specific learning contexts. The interventions were designed to target both motivational and cognitive components of SRL, consistent with the comprehensive model of self-regulation in education outlined above. With this study, we aimed to advance the understanding of how SRL interventions can be optimally designed and implemented using AI, thereby contributing to both theory and practice in educational research. To address these goals, we conducted an RCT with three conditions (i.e., two treatment conditions [*GenAI-supported utility value reflection* and *GenAI-supported cognitive learning strategy prompting*] and a *control condition* [standard ChatGPT use without pedagogical prompting]). We focused on the following research questions (RQs) and corresponding hypotheses (Hs):

RQ1) Did the intervention conditions have the intended effects on SRL processes?

To investigate the intervention's effectiveness, we examined whether the intervention conditions successfully influenced the targeted self-regulatory mechanisms as expected.

H1: Students in the utility value intervention condition show a statistically significantly greater increase in self-reported utility value for daily life than students in the cognitive learning strategy condition.

H2: Students in the cognitive learning strategy condition show a statistically significantly greater increase in self-reported use of the elaboration-based cognitive learning strategy learning-by-explaining than students in the utility value intervention condition.

Additionally, we explored whether students in the cognitive learning strategy condition showed a statistically significantly greater improvement in their performance on the learning-by-explaining task than students in the utility value intervention condition.

RQ2) How does the intervention affect motivational and cognitive learning outcomes?

We investigated whether and to what extent the intervention conditions improved the motivational aspects of learning (i.e., students' maintained interest and effort) and students' learning outcomes compared with the control condition.

H3: Students in the intervention conditions (the utility value and cognitive learning strategy interventions) show a statistically significantly greater increase in maintained interest than students in the control condition.

H4: Students in the intervention conditions (the utility value and cognitive learning strategy interventions) show a statistically significantly greater increase in effort invested in the academic subjects than students in the control condition.

H5: Students in the intervention conditions (the utility value and cognitive learning strategy interventions) show a statistically significantly greater increase in domain-specific knowledge than students in the control condition.

Additionally, we explored the interaction between intervention condition and time, as it remained unclear which intervention condition would yield greater motivational and cognitive learning gains over time.

RQ3) Does the effectiveness of the interventions vary between academic subjects?

We explored whether the interventions' impacts varied by academic subject (English vs. physics). Although we expected the interventions' effects to generalize across academic subjects, given the domain-general nature of SRL processes, we acknowledged that subject-specific differences may influence their impact. For instance, the utility value intervention may be more effective in physics, where connecting abstract concepts (e.g., energy) to students' everyday experiences is more challenging. By contrast, English provides more natural applications (e.g., communication in professional and social contexts, YouTube, websites in English), which may moderate the utility value of the intervention's effects.

With this study, we offer three key innovations in addressing these RQs: We theoretically differentiated between motivational and cognitive-strategy-based GenAI support by using a comprehensive SRL model, we employed transparent and research-based prompt engineering for GenAI tools, and we implemented the intervention in real classroom settings to ensure ecological validity and practical relevance.

Method

To report our study, we followed the CONSORT statement (Cuschieri, 2019).

Sample

An ethics committee approved the study and the data collection (e.g., surveys and testing). It confirmed that the procedures were aligned with ethical research standards with human subjects (date of approval: February 20, 2025, file number: AZ.: A2.5.4–405_hb). In addition, the study was approved by the Ministry of Education, Youth and Sports of the State of Baden-Württemberg.

We employed a convenience sampling approach. Teachers from secondary schools (Gymnasium and Realschule) in the German federal state of Baden-Württemberg were recruited to implement the intervention. The study was introduced to parents during school-organized informational evenings. These informational evenings explained the goals, methodology, theoretical foundation, and data privacy protections of the study. In general, the teachers reported that parental feedback was

positive, and high participation rates across classes indicated strong support. Subsequently, teachers distributed written information and consent forms to the legal guardians and the students, with the vast majority providing consent to participate. These students in Grades 7 to 9 (i.e., lower secondary level) who, along with their legal guardians, provided informed consent and signed the general participant information and privacy policy were included in the study.

The intervention was implemented in regular physics and English lessons. Additional contextual information on classroom implementation, as reported by the participating teachers, is available on our OSF project page (<https://doi.org/10.17605/OSF.IO/Y6KF5>).

Participant recruitment took place from February to April 2025. All students enrolled in the participating classes were eligible and included in the study (i.e., no exclusion criteria were applied). The sample used for the analyses in this study included $N=371$ students (age: $M=13.92$, $SD=1.27$; 54% female; see Table 1 for descriptive statistics for the sample). We conducted a power analysis using G*Power (v 3.1.9.6) before data collection began. The sample size was determined a priori to achieve a power of $\beta-1 = 0.80$, given a small effect size of $f = 0.10$, an alpha level of $\alpha=0.05$, and a mixed ANOVA design with three conditions and two repeated measures. The power analysis indicated that 246 students were needed to meet these criteria. There would thereby be approximately 82 students per condition. However, we collected data from as many students as possible to account for potential attrition and to maximize statistical power and ecological validity. Providing information on age, gender, and migration background was voluntary. 45% of the students reported having a migration background, defined as being born outside Germany or having at least one parent born outside Germany.

In total, we identified 130 students (35%) who dropped out between T1 (pretest) and T2 (posttest), meaning no data were available for them at T2 (utility value intervention condition: $n=37$ [34%], cognitive learning condition: $n=51$ [36%], control condition: $n=42$ [34%]). To examine whether student dropout was systematic, we conducted a regression analysis comparing students who dropped out between T1 and T2 with those who remained in the study on key baseline characteristics (i.e., gender, conscientiousness, openness, dispositional interest, ability self-concept, and growth mindset). The analysis revealed no statistically significant differences between dropouts and non-dropouts on any of these variables, suggesting that attrition was not systematically related to students' initial characteristics.

The two measurements and interventions took place in April and May 2025.

Table 1 Descriptive Statistics for the Sample per Condition

Condition	<i>n</i>	% Female	Drop-outs (<i>n</i>)	% Treatment
GenAI-supported utility value reflection	109	25.7	37	48.6
GenAI-supported cognitive learning strategy prompting	140	35.7	51	47.9
Control condition	122	32.8	42	53.3
Total	371	31.8	130	49.9

Note. *n* = number of participants, Treatment = compliance variable (received treatment) indicating who used the treatment meaningfully

Study Design including Intervention Conditions

Intervention Procedure

We conducted an RCT with a between-subjects design across two academic subjects (i.e., physics and English). Students were randomly assigned to one of three conditions: (a) *GenAI-supported utility value reflection*, (b) *GenAI-supported cognitive learning strategy prompting*, or (c) *control condition* (standard ChatGPT use without pedagogical prompting). Simple randomization was implemented within the IWM Author learning platform (v3.014) using a computer-generated allocation sequence. Students were randomly assigned at the individual level, and the assignments were implemented through the system's backend without disclosing them to teachers or students before the intervention. No blocking or stratification was applied, and slight variations in group sizes resulted from randomization. According to Kang et al. (2008), this approach can be used with large samples ($N > 200$) to achieve overall group balance.

All students interacted with GPT-4o-powered systems. However, system behavior differed across conditions due to differential theory-informed prompting. All prompts were developed through an intensive prompt-engineering process involving multiple iterative feedback loops among the authors and repeated testing phases, following established best practices for effective prompting such as consistent language use, clearly specified task instructions, and a structured knowledge base (e.g., covering the learning tasks, solutions, utility-value intervention principles). All prompts for all conditions and both academic subjects are available in their original German versions and in English translations on our OSF project page. The intervention logic and GPT behavior for each condition were as follows:

GenAI-Supported Utility Value Reflection Drawing on theoretical explanations by Hulleman and Harackiewicz (2009), the GPT encouraged students to reflect on the personal relevance and real-life utility of the learning content. The GPT also connected learning content to students' hobbies, aspirations, or everyday experiences, aiming to increase the perceived task value. An exemplary prompt from the utility-value condition (in English; [Appendix A](#)) and an example dialogue between the chatbot and a student ([Appendix B](#)) are provided. In contrast to, for instance, value-sensitive design approaches (see Nguyen et al., 2025), which aim to proactively embed human values into the design of educational technologies, the prompts adopted an interpretive perspective. The focus was not on redesigning tasks or learning environments to embody specific values, but on supporting students' motivational interpretations of existing instructional activities through GenAI-based scaffolding.

GenAI-Supported Cognitive Learning Strategy Prompting Drawing on theoretical explanations by Lachner et al. (2021), the GPT encouraged students to explain concepts, reflect on their understanding, and self-correct by prompting them to elaborate and asking them metacognitive questions. The GPT steered students toward productive reasoning through Socratic questioning.

Control Condition. The GPT engaged in light conversation about the task, but refrained from offering any pedagogical strategies. It acknowledged the student's input but did not encourage reflection, elaboration, or the use of strategies.

Common Design Features and Implementation Procedure

All GPTs were explicitly instructed not to provide correct answers, even if requested. Although withholding solutions may appear pedagogically inauthentic, prior research indicates that students often either avoid asking for help or request complete solutions prematurely (Aleven et al., 2003; Roll et al., 2014). Experimental evidence further suggests that support emphasizing direct answer provision can reduce generative cognitive engagement and cognitive activation, sometimes described as “metacognitive laziness” (Fan et al., 2025; see also Yan et al., 2025). Accordingly, our design deliberately prioritized cognitive activation by prompting students to articulate reasoning and generate solution ideas rather than receiving answers directly. In doing so, the intervention targeted SRL processes by supporting learners' motivational and strategic engagement, rather than focusing on optimizing performance or providing direct tutoring through answer provision. After completing the tasks and interacting with the chatbot, correct solutions were provided on a subsequent page to allow for corrective feedback and self-correction.

All prompts were adapted to the academic subject (physics or English) but followed the same intervention logic. For instance, the GenAI-supported cognitive learning strategy prompting in English encouraged students to explain grammar rules, whereas the same condition in physics prompted students to explain energy-related principles. The knowledge base for the GPTs consisted of excerpts from the official curriculum relevant to the current unit, ensuring the systems were aware of the content students were expected to learn. In addition, all GPTs had access to the task materials and corresponding solutions to reduce hallucinations.

The study spanned six school sessions (of 45 min each) during regular class time: one pretest session (T1), four consecutive learning sessions, and one posttest session (T2). During the four learning sessions, students worked independently on subject-specific digital tasks (e.g., short open-ended prompts, multiple-choice questions, fill-in-the-blank exercises) via personal computers. In addition to the tasks, students also received explicit instructional input and, where appropriate, worked examples on each learning page. The digital tasks were designed to align with the official curriculum and centered on energy (physics) and job interviews (English, with a focus on grammar, including question formation and verb tenses). Each learning session included three or four structured tasks per academic subject. The materials and tasks were developed by three teachers in close collaboration with two subject-specific education experts for physics and English, ensuring alignment with the official curriculum. Study materials, including instructional content, system prompts, and session documentation, are available on our OSF project page, subject to copyright restrictions.

The intervention was conducted simultaneously across all participating classes and conditions. Most teachers reserved computer rooms or organized personal computers or laptops to ensure that each student worked individually on their computer

using standard German QWERTZ keyboards. Each chatbot appeared under a unified interface design, using the same visual and naming conventions across all conditions. Students interacted with the chatbot by typing their responses on a keyboard; no voice input was used. For each task, the GPT system proactively initiated a conversation 30 s after task onset. The chatbot then engaged in a dynamic dialogue based on the student's input, with the level of interaction depth varying with the student's level of engagement (Fig. 2). As all relevant prompts were embedded in the system, students did not require prior AI experience or prompt-writing skills to interact with the chatbot. For all tasks, all chatbot interfaces displayed the prompt “Gib hier deine Nachricht für ChatGPT ein” (“Enter your message for ChatGPT here”). Messages were submitted by clicking a dedicated “Nachricht abschicken” (“Send message”) button. To ensure synchronized session progression, students could only access the next session after entering a session-specific code announced by the teacher. Similarly, students could proceed to the next task only after interacting with the chatbot interface by entering at least one character into the input field—although submission of the message was not required.

Students logged into the platform using personalized credentials. Teachers provided initial technical support but were instructed to refrain from giving content-related assistance unless students explicitly requested it. Implementation fidelity was supported through standardized instructional procedures and continuous teacher supervision during all learning sessions. All sessions followed the same instructional structure across conditions, including identical instructional content, task formats, session duration, and technical settings; conditions differed only in the chatbot prompts. Teachers' observational data indicated that students varied in the extent to which they interacted with the chatbot, ranging from sustained conversations to more task-focused interactions. In particular, average entry length varied both across conditions and across academic subjects, with longer entries in physics than in English



Fig. 2 Screenshot of learning environment. *Note.* On the left-hand side, students work on tasks (e.g., on the topic of energy in physics lessons), and on the right-hand side, they can interact with the AI assistant through writing

and shorter entries in the cognitive learning condition compared to the utility value condition (see Tables 2 and 3).

Despite a few technical issues during implementation (e.g., login delays, formatting bugs), all planned instructional sessions were conducted as scheduled. Follow-up discussions with teachers after study completion indicated that missing data and student dropout were primarily due to external factors occurring outside the instructional sessions, most notably school-level technical constraints (e.g., temporary Wi-Fi outages during the posttest) and student illness. Teachers consistently reported that dropout was not related to the intervention itself nor to the content or demands of the instruments used. Overall, teachers described the system as stable and classroom engagement as high, despite occasional difficulties in interpreting AI-generated prompts or applying complex formulas in physics lessons.

Data Collection Procedure

For this study, surveys and testing were conducted at two measurement points: T1(pretest) and T2 (posttest). We collected data in schools online using software we developed in-house, called IWM Study (v3.014), which was also used for the intervention.

At T1, which was scheduled before the intervention, participating students were asked about their personality (e.g., conscientiousness), motivational prerequisites (e.g., dispositional interest, academic self-concept), attitudes and experience (e.g., attitudes toward using ChatGPT), and self-regulated learning (i.e., self-reported use of learning strategies). In addition, they were tested on their content knowledge, skill in using the learning-by-explaining strategy, and AI literacy.

At T2, which was scheduled directly after the intervention, the participating students were asked to provide sociodemographic information (e.g., age), personality (i.e., openness), motivational factors (e.g., academic effort), and SRL (i.e., the use of learning strategies). In addition, they were tested on their domain-specific

Table 2 Descriptive statistics of GPT interaction by condition and academic subject

Group	Female <i>n</i> (%)	Treated <i>n</i> (%)	Total words	Average entry length	Number of student entries	Number of meaningful questions	Relative number of meaningful interactions
uv – Physics (<i>n</i> = 79)	23 (46.9%)	42 (53.2%)	1,045 (1191) [0, 5797]	476.34 (225.25) [27.00, 1028.50]	13 (12) [0, 42]	10 (11) [0, 41]	37.80 (32.68) [0.00, 100.00]
uv – English (<i>n</i> = 30)	5 (31.2%)	11 (50.0%)	956 (884) [0, 3630]	452.73 (195.64) [45.17, 768.82]	13 (10) [0, 34]	9 (8) [0, 34]	32.79 (27.59) [0.00, 92.86]
cl – Physics (<i>n</i> = 101)	38 (62.3%)	50 (49.5%)	818 (915) [0, 4111]	333.16 (183.04) [78.57, 1679.67]	15 (14) [0, 68]	12 (13) [0, 61]	34.58 (28.41) [0.00, 100.00]
cl – English (<i>n</i> = 39)	12 (70.6%)	17 (73.9%)	811 (546) [0, 1951]	285.46 (57.26) [182.78, 420.11]	18 (11) [0, 44]	14 (9) [0, 37]	46.27 (29.68) [0.00, 100.00]
co – Physics (<i>n</i> = 90)	35 (57.4%)	51 (56.7%)	1,089 (1161) [0, 5149]	498.51 (273.05) [25.00, 1608.00]	14 (11) [0, 50]	11 (10) [0, 46]	38.45 (30.75) [0.00, 100.00]
co – English (<i>n</i> = 32)	5 (31.2%)	14 (56.0%)	971 (1084) [0, 4541]	312.84 (121.50) [142.80, 637.90]	18 (16) [0, 57]	13 (15) [0, 46]	40.29 (34.43) [0.00, 100.00]

Values are *M* (*SD*) [*Min*, *Max*] for continuous/count variables and *n* (%) for binary variables. uv = GenAI-supported utility value reflection, cl = GenAI-supported cognitive learning strategy prompting, co = control condition

Table 3 Differences of GPT interaction by condition and academic subject

Variable	Condition <i>p</i>	Academic Subject <i>p</i>
Total words	0.147	0.620
Average entry length	< 0.001	< 0.001
Number of student entries	0.240	0.184
Number of meaningful questions	0.166	0.503
Relative number of meaningful interactions	0.833	0.457
Received treatment	0.833	0.358

Condition *p* values: one-way ANOVA (continuous/count) or χ^2 /Fisher (binary). Subject *p* values: *t* tests (continuous/count) or χ^2 /Fisher (binary). Gender (GS) was not tested

knowledge, skill in using the learning-by-explaining strategy, and cognitive ability. At both measurement points, the students had 45 min in one lesson to complete the questionnaire and the test tasks, a sufficient amount of time for all students.

During the intervention, a brief (i.e., 23-item) self-assessment (e.g., regarding mental effort) was conducted online immediately after each session. Moreover, log data on students' utilization of the online learning system IWM Study (v. 3.014) were assessed. This study does not include data from the short self-assessment or the log data. The RCT was completed as planned after T2.

The data used in this study are available on our OSF project page.

Measures

Whenever we used scales, we utilized manifest scale scores (i.e., arithmetic means). All dependent variables were administered at T1 and T2 to capture changes over time. The measures were situated within the framework that is presented in Fig. 1.

Dependent Variables

Utility Value of the Learning Content Our primary outcome for the utility value intervention condition was students' perceived utility value of the learning content for daily life, which was assessed using a self-report scale comprising three items (e.g., "Understanding < topic > has many advantages in my daily life"; translated from Gaspard et al., 2015). Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*). The original items were adapted, with the first item modified in the German version to refer specifically to media education. The internal consistency in our data was high (physics: T1: $\alpha = 0.84$; T2: $\alpha = 0.89$; English: T1: $\alpha = 0.78$; T2: $\alpha = 0.91$) on the basis of Taber et al.'s (2018) criteria.

Elaboration-Based Strategy Our primary outcome for the cognitive learning condition was students' self-reported use of the elaboration-based cognitive learning strategy, which was assessed with a three-item self-report scale (e.g., "I try to relate new concepts or theories to concepts and theories I already know"; translated from Klingensieck, 2018). Responses were rated on a 5-point scale ranging from 1 (*not true at all*) to 5 (*exactly true*). The internal consistency in our data was acceptable to high (T1: $\alpha = 0.69$; T2: $\alpha = 0.80$) on the basis of Taber et al.'s (2018) criteria.

Application of the Learning-by-Explaining Strategy To assess the use of the elaboration-based learning strategy, learning-by-explaining, students completed an open-ended task at T1 and T2 (see Fig. 3 for an example from the subject of English). In this task, students read a short text on a recently covered classroom topic and were instructed to explain its content in writing as if they were explaining it to a classmate. The T1 and T2 tasks differed in content but followed the same structure. This approach assessed a student's ability to apply the learning-by-explaining strategy. The teachers who developed the tasks also rated the depth and accuracy of the explanations. To reduce bias, they were blinded to students' identities (using anonymous codes) and experimental conditions. No formal interrater agreement was calculated, as the responses were not double-coded.

Maintained Interest Students' maintained interest was assessed with a self-report subscale comprising four items (e.g., "< topic> fascinates me."; translated and adapted from Linnenbrink-Garcia et al., 2010). Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*). The internal consistency in our data was high (T1: $\alpha = 0.91$; T2: $\alpha = 0.95$) on the basis of Taber et al.'s (2018) criteria.

Effort Investment Students' self-reported effort investment in the academic subject (effort) was assessed with a five-item scale (e.g., "I put a lot of effort into understanding the subject matter thoroughly"; translated from Jaekel et al., 2020). Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*). The internal consistency in our data was high (physics: T1: $\alpha = 0.86$; T2: $\alpha = 0.88$; English: T1: $\alpha = 0.86$; T2: $\alpha = 0.90$) on the basis of Taber et al.'s (2018) criteria.

Domain-Specific Knowledge. Students' domain-specific knowledge was assessed using three subject-specific test items (e.g., single-choice, matching, multiple-choice, fill-in-the-blank) tailored to the content covered in the interventions. The same items were administered at T1 and T2 to evaluate students' learning progress. To ensure ecological validity, all items were embedded in the lesson content and aligned with typical classroom assessments in regular physics and English curricula. Given the constraints of classroom time, the number of test items was intentionally limited to allow a time-efficient, context-appropriate assessment. Content validity was established through a structured, iterative expert-informed development process. For each academic subject, experienced in-service teachers developed curriculum-aligned test items and instructional materials, which were subsequently reviewed, discussed, and refined in multiple feedback rounds together with subject-specific educational researchers (professors in subject-matter education) to ensure curricular alignment, conceptual accuracy, and instructional relevance. The items were rated by the same subject teachers who had developed the tests, tasks, and instructional materials. An example item from physics asked students to identify the correct energy flow diagram in a familiar everyday scenario (see Fig. 4). All items were designed to assess students' conceptual understanding and basic application of subject knowledge. To reduce bias, all scoring was conducted using anonymized student codes, and teachers were blind to the experimental condition. To ensure comparability across academic

German Version

Erklärung von Fachinhalten

Unten siehst du drei englische Sätze, in denen eine Frage gestellt wird.

Stell dir vor, du willst einem Mitschüler oder einer Mitschülerin erklären, wie diese Frage funktioniert.

Du musst keine Fachbegriffe verwenden. Erkläre einfach, was dir an der Satzstellung oder dem Aufbau auffällt.

Nutze gerne diese Satzanfänge:

Ich glaube, das ist eine Frage, weil...

Der Satz beginnt mit...

Das Wort ... zeigt, dass...

So eine Frage kenne ich aus dem Alltag...

Bitte erkläre die Struktur der Fragen kurz schriftlich so, dass eine Freundin oder ein Freund von dir aus der Klasse die Fragebildung versteht.

Beispiel 1:

How long have you lived here?

Deine Erklärung:

English Version

Explaining Subject Content

Below, you will see three English sentences that are questions.

Imagine you want to explain to a classmate how this question works.

You don't need to use technical terms. Just explain simply what you notice about the word order or the structure.

You may use the following sentence starters:

I think this is a question because...

The sentence starts with...

The word ... shows that...

I know this kind of question from everyday life...

Please explain the structure of the questions briefly in writing so that a friend of yours from class can understand how to form questions.

Example 1:

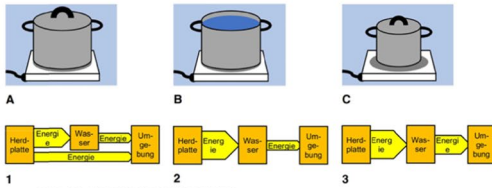
How long have you lived here?

Your explanation:

Fig. 3 Example of a test item used to assess the application of learning-by-explaining in English. *Note.* The German version was used in the study

German Version

Durch die Dicke der Energie-Pfeile kann man bei einem Energieflussdiagramm darstellen, ob viel oder wenig Energie übertragen wird. Wenn Energie auf verschiedene Gegenstände übertragen wird, gibt es Verzweigungen.



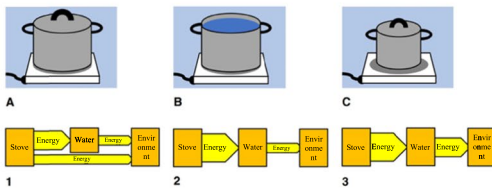
Source: C.J. Pandall – ZPG BNT (CC) BY-NC-SA 3.0 DE

Ordne jeder Koch-Situation A/B/C ein passendes Energieflussdiagramm 1/2/3 zu.

Koch-Situation A
Koch-Situation B
Koch-Situation C
Energieflussdiagramm 1
Energieflussdiagramm 2
Energieflussdiagramm 3

English Version

The thickness of the energy arrows in an energy flow diagram indicates whether a lot or little energy is transferred. When energy is transferred to different objects, branches appear.



Source: C.J. Pandall – ZPG BNT (CC) BY-NC-SA 3.0 DE

Match each cooking situation A/B/C to a suitable energy flow diagram 1/2/3.

Cooking Situation A
Cooking Situation B
Cooking Situation C
Energy Flow Diagram 1
Energy Flow Diagram 2

Fig. 4 Example test item in physics. *Note.* For this task, students could receive a minimum of 0 and a maximum of 3 points. The German version was used in the study

subjects, raw test scores from the physics and English knowledge tests were standardized (z scores) prior to analysis. This standardization was performed separately for each subject. The resulting z scores reflect students' relative performance within each subject domain.

Independent Variable

Students were randomly assigned to one of three experimental conditions, which will be indicated by a *condition* variable in the data set (0=*control condition*, 1=*utility value condition*, 2=*cognitive learning condition*).

Control Variables

We considered three variables as potential control variables in robustness checks: (a) academic subject (0=*physics*, 1=*English*), (b) students' self-reported gender, assessed using a single-item question (0=*male*, 1=*female*, 2=*diverse*; responses indicating diverse were recoded as missing because the number of students selecting this option was too small to yield interpretable or reliable statistical estimates), and (c) students' self-reported migration background, assessed via three items capturing the country of birth of students and their parents. Students were classified as having

no migration background (=0) if they and both parents were born in Germany; all other constellations were classified as having a migration background (=1). All variables were assessed at T1.

Following the What Works Clearinghouse (WWC) Standards (2022), we examined baseline equivalence for these variables. Baseline differences were evaluated using standardized effect sizes (ES), with $0.00 \leq |ES| \leq 0.05$ indicating baseline equivalence, $0.05 < |ES| \leq 0.25$ requiring statistical adjustment, and $|ES| > 0.25$ indicating a failure to meet baseline equivalence. Variables were included as covariates in robustness checks (i.e., all variables except AI literacy were included as covariates) when baseline nonequivalence was detected (see Table 4). Importantly, variables were included as covariates only when baseline nonequivalence exceeded the threshold of $|ES| > 0.05$ across all pairwise condition comparisons. Moreover, in cases where standardized differences slightly exceeded the 0.25 threshold, we also included the respective variables as covariates to account for estimation uncertainty, given that the WWC guidelines were developed primarily for large-scale RCTs with standardized assessments and that effect size estimates are inherently less stable in smaller samples.

Table 4 Baseline differences in potential control variables

Variable	Condition	N	M	SD	Baseline difference Cohen's <i>d/p</i>
Cognitive ability	uv	109	5.21	6.09	uv – cl: $d = 0.26, p = .046$
	cl	140	3.79	4.71	uv – co: $d = 0.17, p = .190$
	co	122	4.25	4.89	cl – co: $d = -0.10, p = .440$
Conscientiousness	uv	104	3.11	0.72	uv – cl: $d = 0.10, p = .440$
	cl	133	3.03	0.72	uv – co: $d = -0.09, p = .510$
	co	119	3.17	0.74	cl – co: $d = -0.19, p = .140$
Openness	uv	70	3.27	0.73	uv – cl: $d = 0.15, p = .340$
	cl	88	3.15	0.80	uv – co: $d = -0.07, p = .690$
	co	80	3.31	0.68	cl – co: $d = -0.22, p = .160$
Academic self-concept	uv	104	2.62	0.71	uv – cl: $d = 0.06, p = .630$
	cl	134	2.58	0.70	uv – co: $d = -0.09, p = .530$
	co	119	2.68	0.69	cl – co: $d = -0.15, p = .230$
Dispositional interest	uv	104	2.42	0.85	uv – cl: $d = -0.05, p = .730$
	cl	133	2.45	0.86	uv – co: $d = -0.28, p = .037$
	co	119	2.65	0.79	cl – co: $d = -0.23, p = .065$
Growth mindset	uv	104	2.76	0.63	uv – cl: $d = -0.22, p = .093$
	cl	134	2.89	0.57	uv – co: $d = -0.38, p = .005$
	co	119	2.99	0.60	cl – co: $d = -0.17, p = .170$
AI literacy	uv	109	2.28	1.45	uv – cl: $d = -0.04, p = .770$
	cl	140	2.33	1.41	uv – co: $d = -0.08, p = .530$
	co	122	2.39	1.41	cl – co: $d = -0.05, p = .710$

Self-concept and dispositional interest are reported across both subjects here. On our OSF project page, the output files also include subject-specific differences between the conditions. uv = GenAI-supported utility value reflection, cl = GenAI-supported cognitive learning strategy prompting, co = control condition

Cognitive Ability Students' cognitive ability was assessed at T2 using the KFT 4–12+R cognitive ability test (Heller & Perleth, 2000), which uses a figure analogy task to measure reasoning skills. The test consisted of the first 20 multiple-choice items (reduced from the original 25 items; e.g., ‘Which of the five figures on the right completes the given analogy?’), each with five response options. Responses were coded dichotomously (1 = *correct*, 0 = *incorrect*), and the sum score was used to represent cognitive ability. Internal consistency was estimated with KR-20, which is suitable for dichotomously scored items ($KR-20 = 0.87$) and is in line with previous reports ($0.80 \leq \alpha \leq 0.90$; Heller & Perleth, 2000).

Conscientiousness Students' conscientiousness was assessed at T1 via self-report using a four-item scale (Danner et al., 2016). An example item was “I stick with a task until it is completed.” Responses were rated on a 5-point scale ranging from 1 (*does not apply at all*) to 5 (*fully applies*), with two reverse-coded items. The internal consistency in our data was moderate ($\alpha = 0.59$) on the basis of Taber et al.'s (2018) criteria.

Openness Students' openness was assessed at T2 via a four-item self-report scale (Danner et al., 2016). An example item was “I enjoy thinking deeply about complex things and understanding them.” Responses were rated on a 5-point scale ranging from 1 (*does not apply at all*) to 5 (*fully applies*), with two reverse-coded items. The internal consistency in our data was not satisfactory ($\alpha = 0.39$) on the basis of Taber et al.'s (2018) criteria.

Academic Self-Concept. Students' academic self-concept was assessed at T1 with a four-item self-report scale (Gaspard et al., 2016). An example item was “I am good at [physics/English].” Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*), with two reverse-coded items. The internal consistency in our data was high (physics: $\alpha = 0.78$; English: $\alpha = 0.85$).

Dispositional Interest Students' dispositional interest in the academic subject was assessed at T1 with a four-item self-report scale (Gaspard et al., 2017). An example item was “I enjoy engaging with topics in [English/physics].” Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*). The internal consistency in our data was high (physics: $\alpha = 0.94$; English: $\alpha = 0.93$) on the basis of Taber et al.'s (2018) criteria.

Growth Mindset Students' growth mindset was assessed at T1 with a three-item self-report scale (Rammstedt et al., 2024). An example item was “You can learn new things, but you can't really change your basic intelligence.” All items were reverse-coded. Responses were rated on a 4-point scale ranging from 1 (*not true at all*) to 4 (*exactly true*). The internal consistency in our data was acceptable ($\alpha = 0.70$) on the basis of Taber et al.'s (2018) criteria.

AI Literacy Students' AI literacy was assessed at T1 with eight single-choice items selected from the AI literacy test developed by Hornberger et al. (2023). Each item

had four answer options, of which only one was correct. The items covered different dimensions of AI literacy, including recognizing AI, understanding intelligence, identifying AI strengths and weaknesses, and considering ethical implications. An example item was “Which of the following areas typically involves AI applications?” with the correct response option (1) Detecting credit card fraud. Responses were coded dichotomously (0 = *incorrect*, 1 = *correct*), and the sum score was used to represent AI literacy. The internal consistency of our selected items was not satisfactory ($KR-20 = 0.23$), whereas previous findings for the whole AI literacy test showed high internal consistency (i.e., $\alpha = 0.82$; Hornberger et al., 2023).

Compliance Variable used for Robustness Checks

We included the variable *treatment received* to estimate the Complier Average Causal Effect (CACE; Nagengast et al., 2018) using Instrumental Variable Regression (Angrist et al., 1996; Angrist & Pischke, 2015). This variable was derived from students’ actual meaningful use of the GPT systems. First, we counted the number of meaningful interactions students had with the GPTs. An interaction was classified as meaningful if the students exchanged at least one task-related and non-trivial message (i.e., messages that went beyond brief or generic utterances such as greetings, acknowledgments, or off-task remarks such as “hey” or “I love you” and instead contained substantive, task-focused content such as questions, elaborations, or clarifications related to the task) per task. Second, we counted the number of possible interactions with the GPTs across but not within tasks (i.e., the number of tasks is the maximum). Third, we calculated the relative number of meaningful interactions each student participated in, expressed as a percentage of the total expected number of conversations, because the number of tasks varied between physics and English lessons. Finally, we created the *treatment received* variable as a binary indicator (0 = *noncompliers*, i.e., *students who did not use the GPT meaningfully at least once per task*; 1 = *compliers*, i.e., *students who used the GPT meaningfully at least once per task*). This variable, along with the *assigned treatment condition* variable, distinguished between compliers and noncompliers in the analysis.

An overview of all variables constructed to capture students’ interactions with the GPT system is provided in Table A in the Appendix. An overview of the descriptive statistics for these variables is presented in Table 2a and Table 2b. A codebook and descriptions of the items and scales are available online on our OSF project page.

Statistical Data Analyses

We computed all analyses in *R* (v4.5.0, R Core Team, 2025) and utilized *RStudio* (v2025.5.1.513, Posit team, 2025).

To analyze our 3 (between-subjects factor: *condition*) $\times$ 2 (within-subjects factor: *time*) mixed design, we used linear mixed-effects models (LMMs) with the *lmer()* function from the *R* package *lme4* (v1.1.37, Bates et al., 2015). In all models, we included fixed effects for the variables *condition* (i.e., utility value, cognitive learning strategy, control), *time* (i.e., T1 vs. T2), and their interaction to test our hypotheses. We included random intercepts for participants to account for individual differences

in repeated measures. For exploratory analyses, we included the variable *subject* (i.e., English vs. physics) as an additional predictor and examined three-way interactions. We used effect coding for categorical predictors to facilitate the interpretation of main effects and interactions.

More specifically, to answer RQ1 and test H1 and H2, we tested whether the intervention successfully influenced SRL mechanisms as expected. To get the *F* tests of the interaction, we conducted a multiparameter test (i.e., an ANOVA-like test) using model comparison, implemented by the *anova()* function in *R*. Next, we computed post hoc contrasts using the function *emmeans()* from the *R* package *emmeans* (v1.11.1, Lenth, 2024) for pairwise comparisons of conditions. Additionally, we explored whether the cognitive learning strategy condition led to a greater increase in the application of the elaboration-based strategy, *learning-by-explaining*.

To answer RQ2 and test H3-H5, we performed the same analyses used to test H1 and H2, using students' maintained interest, effort, and learning outcomes as the dependent variables. We investigated whether the intervention conditions improved the dependent variables relative to the control condition. Additionally, we explored whether cognitive and motivational learning gains changed over time across intervention conditions.

To answer RQ3, we explored whether the intervention effects analyzed to answer RQ2 varied across academic subjects (i.e., English vs. physics).

Handling of Missing Data

Across all indicators at both measurement points, 21% of the numerical data were missing (based on 83,104 entries in total). When considering only the scale scores, the missing-data rate was 16%. A closer look at the percentage of missing values for all indicators of the dependent variable at T1 (5%) compared with T2 (36%) suggested that a substantial portion of the missing data could be attributed to student dropout between the two measurement points. This interpretation is further supported by the observation that, when restricting the analysis to the 241 students who did not drop out of the study, missingness was minimal at both measurement points (T1: 0.14%; T2: 1.73%). If any individual scale items were missing, we calculated the scale mean using the available values (i.e., *na.rm=TRUE* in *R*). Furthermore, we did not take any measures to estimate or impute missing values in the primary analyses. However, the first robustness check involved rerunning the analyses with imputed data.

Robustness Checks

To test the robustness of our findings, we conducted several robustness checks. As a first robustness check, we repeated the analyses using imputed data to ensure our findings were not biased by missing data. We conducted multiple imputation using the *R* package *mice* (v3.18.0, van Buuren & Groothuis-Oudshoorn, 2011), generating 50 imputed data sets. We then tested the same multilevel models within each imputed data set using the *with()* function from the *mice* package and pooled the resulting estimates using *pool()* as implemented in the package (i.e., following Rubin's rules) to account for the variability introduced by imputation.

As a second robustness check, we repeated the analyses using data from students who did not drop out.

As a third robustness check, to ensure that our findings were not biased by potential preexisting differences between conditions, we re-estimated the multilevel models while including additional control variables. We included only control variables that exhibited baseline nonequivalence on the basis of the WWC standards (2022). In the first step, we examined whether the control variables differed between the conditions. Depending on the measurement scale of the variables, we used appropriate statistical tests (i.e., *t* tests, one-way ANOVA, Pearson's χ^2 test). In accordance with WWC guidelines, we considered a variable to show baseline nonequivalence if the standardized mean difference (Cohen's *d*) exceeded 0.05, regardless of statistical significance. In a second step, we re-estimated the models, including only the control variables that exhibited baseline nonequivalence.

Finally, we estimated a Complier Average Causal Effect (CACE) to assess the intervention's effect on students who actively participated (Nagengast et al., 2018). Whereas the primary analyses (e.g., ANOVAs) followed an intention-to-treat (ITT) approach, which included all students assigned to the conditions, CACE specifically estimated the intervention effect in those who actually received it. This approach assumes that random assignment functions as an instrumental variable for participation. We implemented CACE estimation using instrumental variable regression (IVR; Angrist & Pischke, 2015), specifically two-stage least squares using the AER package (v1.2.14, Kleibers & Zeileis, 2008), and extended it to a multilevel mixed-effects model to account for hierarchical data structures.

We controlled for false discovery rates in all our models by applying the Benjamini and Hochberg (1995) correction per model. The significance level was set to $\alpha=0.05$ for all analyses, regardless of whether the hypotheses were directional or nondirectional. This study, including the data analysis approaches, was preregistered at OSF: <https://doi.org/10.17605/OSF.IO/VBSQ4>. The full reproducible code (i.e., we used R Markdown and the *renv* package) and data are online on our OSF project page.

Results

Preliminary Results

To provide an overview of all central independent, dependent, and control variables, we report descriptive statistics (i.e., *M*, *SD*, *Min*, *Max*; Table 5), the results of baseline equivalence testing for potential control variables (Table 4), and the correlations between the scales (Table 6).

In accordance with WWC standards, the results indicated that, except for the test of AI literacy, all potential control variables should be included in our robustness checks. Although most variables did not show statistically significant baseline differences across the three conditions at T1, their inclusion as controls was still justified to account for potential confounding.

Table 5 Descriptive statistics for all variables and scales used in this study

Variable/Scale	N	Min	Max	M	SD
Age ^a	231	12	18	13.92	1.27
% Gender ^b	220	0	1	0.54	0.50
% Compliance (i.e., used treatment meaningfully)	340	0	100	37.45	30.50
% Tested knowledge at T1 ^c (across academic subjects)	364	7.69	100	59.77	16.80
% Tested knowledge at T2 ^c (across academic subjects)	239	7.69	100	58.56	18.23
% Tested knowledge at T1 ^c (physics)	267	20	86.67	54.22	13.12
% Tested knowledge at T2 ^c (physics)	183	10	80	52.57	12.27
% Tested knowledge at T1 ^c (English)	97	7.69	100	75.06	16.39
% Tested knowledge at T2 ^c (English)	56	7.69	100	78.16	20.72
Application of learning-by-explaining at T1 ^d (across academic subjects)	352	0	6	1.90	1.95
Application of learning-by-explaining at T2 ^d (across academic subjects)	241	0	6	1.67	1.87
Application of learning-by-explaining at T1 ^d (physics)	261	0	6	1.81	1.80
Application of learning-by-explaining at T2 ^d (physics)	185	0	6	1.70	1.87
Application of learning-by-explaining at T1 ^d (English)	91	0	6	2.15	2.30
Application of learning-by-explaining at T2 ^d (English)	56	0	6	1.57	1.90
Utility value at T1 ^e (across academic subjects)	356	1	4	2.60	0.72
Utility value at T2 ^e (across academic subjects)	238	1	4	2.55	0.80
Utility value at T1 ^e (physics)	263	1	4	2.62	0.74
Utility value at T2 ^e (physics)	182	1	4	2.58	0.81
Utility value at T1 ^e (English)	93	1	4	2.52	0.63
Utility value at T2 ^e (English)	56	1	4	2.52	0.63
Self-reported elaboration-based strategy at T1 ^f	344	1	5	3.19	0.82
Self-reported elaboration-based strategy at T2 ^f	231	1	5	3.15	0.91
Effort investment at T1 ^e (across academic subjects)	353	1	4	2.72	0.67
Effort investment at T2 ^e (across academic subjects)	239	1	4	2.77	0.69
Effort investment at T1 ^e (physics)	262	1	4	2.70	0.69
Effort investment at T2 ^e (physics)	183	1	4	2.76	0.68
Effort investment at T1 ^e (English)	91	1.2	4	2.78	0.63
Effort investment at T2 ^e (English)	56	1	4	2.80	0.72
Maintained interest at T1 ^e	354	1	4	2.35	0.76
Maintained interest at T2 ^e	237	1	4	2.32	0.84
Cognitive ability ^g	371	0	20	4.36	5.23
Conscientiousness ^g	356	1	5	3.10	0.73
Openness ^g	238	1.67	5	3.24	0.74
Academic self-concept ^e (across academic subjects)	357	1	4	2.63	0.70
Academic self-concept ^e (physics)	263	1	4	2.56	0.66
Academic self-concept ^e (English)	94	1	4	2.81	0.77
Dispositional interest ^e (across academic subjects)	356	1	4	2.51	0.83
Dispositional interest ^e (physics)	263	1	4	2.45	0.83
Dispositional interest ^e (English)	93	1	4	2.66	0.84
Growth mindset ^e	357	1	4	2.89	0.61
AI literacy ^h	371	0	6	2.33	1.41

On our OSF project page, we provide a codebook in which all indicator values across subjects are documented and accessible.

^a in years, ^b 0 = male, 1 = female, ^c percentage of correct answers, ^d 0 to 6 points, ^e 1 = not true at all to 4 = exactly true, ^f 1 = not true at all to 5 = exactly true, ^g 1 = does not apply at all to 5 = fully applies, ^h 0 to 20 points, ⁱ 0 to 8 points

Table 6 Correlations between all scales used in the study

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relative number of meaningful interactions (1)	-																			
Tested knowledge at T1 (2) ^a	0.12, $p = .034$	-																		
Tested knowledge at T2 (3) ^a	0.14, $p = .032$	0.35, $p < .001$	-																	
Application of learning-by-explaining at T1 (4) ^a	0.24, $p < .001$	0.25, $p < .001$	0.16, $p = .016$	-																
Application of learning-by-explaining at T2 (5) ^a	0.17, $p = .011$	0.21, $p = .001$	0.13, $p = .052$	0.56, $p < .001$	-															
Cognitive ability (6)	0.35, $p < .001$	0.17, $p = .001$	0.14, $p = .032$	0.25, $p < .001$	0.41, $p < .001$	-														
AI literacy (7)	0.22, $p < .001$	0.23, $p < .001$	0.11, $p < .001$	0.28, $p < .001$	0.32, $p < .001$	0.34, $p < .001$	-													
Utility value at T1 (8)	-0.05, $p = .371$	-0.03, $p = .623$	-0.00, $p = .963$	0.01, $p = .792$	-0.13, $p = .044$	-0.13, $p = .014$	-0.13, $p = .004$	-0.15, $p = .004$	-											
Utility value at T2 (9)	0.05, $p = .413$	0.03, $p = .677$	0.10, $p = .118$	0.02, $p = .714$	-0.01, $p = .911$	0.01, $p = .930$	-0.01, $p = .910$	0.50, $p < .001$	-											

Table 6 (continued)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Self-reported elaboration-based strategy at T1 (10)	0.05 <i>p</i> =.332	0.13 <i>p</i> =.017	0.10 <i>p</i> =.114	0.19 <i>p</i> <.001	0.15 <i>p</i> =.023	0.12 <i>p</i> =.025	0.22 <i>p</i> <.001	0.08 <i>p</i> =.123	0.18 <i>p</i> =.006	-										
Self-reported elaboration-based strategy at T2 (11)	0.02 <i>p</i> =.757	0.06 <i>p</i> =.348	0.09 <i>p</i> =.157	0.12 <i>p</i> =.068	0.15 <i>p</i> =.025	0.15 <i>p</i> =.026	0.15 <i>p</i> =.024	0.13 <i>p</i> =.048	0.32 <i>p</i> <.001	0.38 <i>p</i> <.001	-									
Effort investment at T1 (12)	0.05 <i>p</i> =.354	0.10 <i>p</i> =.053	0.18 <i>p</i> =.005	0.20 <i>p</i> <.001	0.12 <i>p</i> =.053	0.02 <i>p</i> =.658	-0.08 <i>p</i> =.124	0.34 <i>p</i> <.001	0.28 <i>p</i> <.001	0.31 <i>p</i> <.001	0.29 <i>p</i> <.001	-								
Effort investment at T2 (13)	0.11 <i>p</i> =.084	0.08 <i>p</i> =.202	0.13 <i>p</i> =.041	0.13 <i>p</i> =.038	0.12 <i>p</i> =.061	0.04 <i>p</i> =.547	0.02 <i>p</i> =.708	0.31 <i>p</i> <.001	0.49 <i>p</i> <.001	0.23 <i>p</i> <.001	0.35 <i>p</i> <.001	0.64 <i>p</i> <.001	-							
Maintained interest at T1 (14) ^a	0.02 <i>p</i> =.745	-0.03 <i>p</i> =.576	0.02 <i>p</i> =.783	0.05 <i>p</i> =.378	0.04 <i>p</i> =.553	0.01 <i>p</i> =.882	-0.05 <i>p</i> =.338	0.48 <i>p</i> <.001	0.49 <i>p</i> <.001	0.14 <i>p</i> =.008	0.31 <i>p</i> <.001	0.51 <i>p</i> <.001	0.48 <i>p</i> <.001	-						
Maintained interest at T2 (15) ^a	-0.03 <i>p</i> =.629	0.08 <i>p</i> =.232	0.07 <i>p</i> =.294	0.03 <i>p</i> =.677	0.01 <i>p</i> =.937	0.01 <i>p</i> =.881	0.02 <i>p</i> =.805	0.44 <i>p</i> <.001	0.70 <i>p</i> <.001	0.18 <i>p</i> =.005	0.30 <i>p</i> <.001	0.35 <i>p</i> <.001	0.53 <i>p</i> <.001	0.63 <i>p</i> <.001	-					
Conscientiousness (16)	0.01 <i>p</i> =.828	0.12 <i>p</i> =.019	0.18 <i>p</i> =.007	0.01 <i>p</i> =.796	-0.01 <i>p</i> =.796	-0.05 <i>p</i> =.346	0.01 <i>p</i> =.916	0.17 <i>p</i> =.001	0.19 <i>p</i> =.003	0.24 <i>p</i> <.001	0.22 <i>p</i> =.001	0.42 <i>p</i> <.001	0.38 <i>p</i> <.001	0.24 <i>p</i> <.001	0.27 <i>p</i> <.001	-				
Openness (17)	0.09 <i>p</i> =.179	0.14 <i>p</i> =.003	0.26 <i>p</i> <.001	0.23 <i>p</i> <.001	0.21 <i>p</i> =.001	0.28 <i>p</i> <.001	0.27 <i>p</i> <.001	0.08 <i>p</i> =.233	0.19 <i>p</i> =.003	0.36 <i>p</i> <.001	0.35 <i>p</i> <.001	0.20 <i>p</i> =.002	0.31 <i>p</i> <.001	0.17 <i>p</i> =.007	0.23 <i>p</i> <.001	0.24 <i>p</i> <.001	-			
Academic self-concept (18) ^a	0.06 <i>p</i> =.298	0.17 <i>p</i> =.001	0.23 <i>p</i> <.001	0.12 <i>p</i> =.001	0.14 <i>p</i> =.001	0.14 <i>p</i> <.001	0.04 <i>p</i> <.001	0.10 <i>p</i> =.233	0.18 <i>p</i> =.003	0.33 <i>p</i> <.001	0.34 <i>p</i> <.001	0.46 <i>p</i> =.002	0.41 <i>p</i> <.001	0.31 <i>p</i> =.007	0.21 <i>p</i> <.001	0.26 <i>p</i> <.001	0.27 <i>p</i> <.001	-		

Table 6 (continued)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Dispositional interest (19) ^a	0.06 <i>p</i> = .295	0.02 <i>p</i> = .742	0.14 <i>p</i> = .036	0.13 <i>p</i> = .015	0.12 <i>p</i> = .064	0.11 <i>p</i> = .032	-0.05 <i>p</i> = .370	0.28 <i>p</i> = .001	0.31 <i>p</i> = .001	0.24 <i>p</i> = .001	0.35 <i>p</i> = .001	0.53 <i>p</i> = .001	0.48 <i>p</i> = .001	0.51 <i>p</i> = .001	0.38 <i>p</i> = .001	0.19 <i>p</i> = .001	0.31 <i>p</i> = .001	0.64 <i>p</i> = .001	-	-
Growth mind-set (20)	0.07 <i>p</i> = .180	0.04 <i>p</i> = .465	-0.07 <i>p</i> = .295	0.08 <i>p</i> = .126	0.05 <i>p</i> = .426	-0.01 <i>p</i> = .888	0.06 <i>p</i> = .264	0.05 <i>p</i> = .302	0.16 <i>p</i> = .015	0.11 <i>p</i> = .035	0.13 <i>p</i> = .049	0.03 <i>p</i> = .636	0.25 <i>p</i> = .001	0.07 <i>p</i> = .196	0.17 <i>p</i> = .010	0.15 <i>p</i> = .005	0.07 <i>p</i> = .317	0.04 <i>p</i> = .436	0.06 <i>p</i> = .233	-

^a Across academic subjects

As shown in Table 6, as expected, several of the outcome variables were substantially correlated. For instance, maintained interest was positively associated with effort investment (T1: $r = .51, p < .001$; T2: $r = .53, p < .001$). Moreover, the stabilities between T1 and T2 for these two dependent variables were as expected (e.g., maintained interest: $r = .63, p < .001$; effort investment: $r = .64, p < .001$).

Intervention Effects on Students' Self-Regulated Learning (RQ1)

To test whether the utility value intervention led to greater increases in students' *perceived utility value* (H1) or *self-reported elaboration-based strategy use* (H2) compared with the cognitive learning strategy condition, we estimated a linear mixed-effects model with time (T1, T2) and condition (utility value, cognitive learning strategy, control) as fixed effects, and a random intercept for participants. Additionally, we explored students' application of the elaboration-based strategy (i.e., learning-by-explaining).

Perceived Utility Value

Regarding *perceived utility value* (H1), the interaction between time and condition was statistically significant, $\chi^2(2) = 19.05, p < .001$, indicating that utility value changed differently across conditions (Fig. 5). Specifically, students in the GenAI-supported

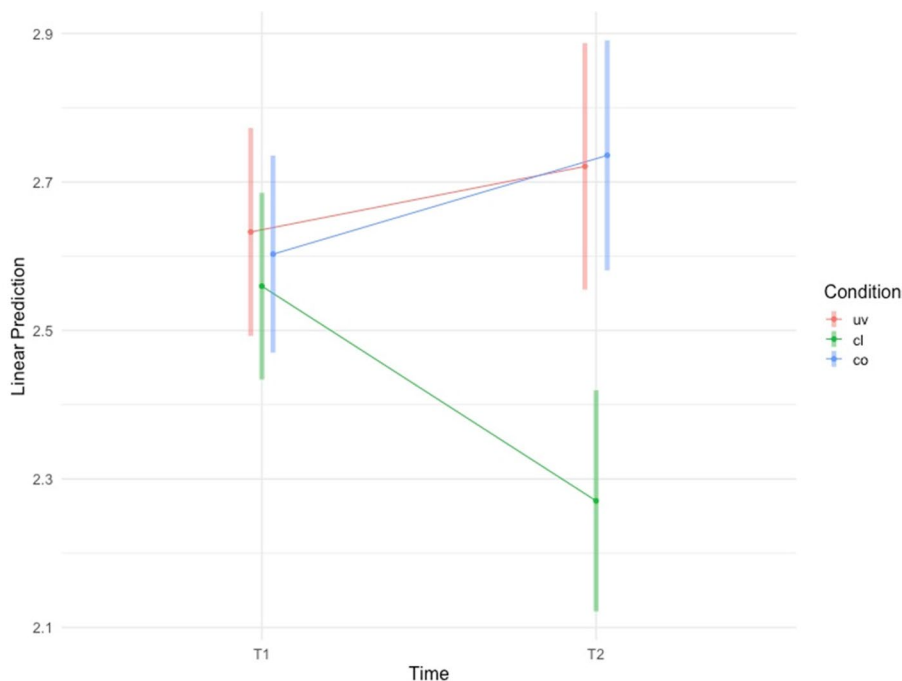


Fig. 5 Estimated marginal means of self-reported utility value by intervention condition and time. Note. Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

utility value reflection condition showed a statistically significantly more favorable trajectory compared with those in the GenAI-supported cognitive learning strategy prompting condition ($\beta = -0.50$, $SE = 0.11$, $p < .001$). However, this difference was not observed between the GenAI-supported utility value reflection condition and the control condition. Within-group comparisons revealed a statistically significant decrease in utility value in the GenAI-supported cognitive learning strategy prompting condition ($\Delta M = 0.30$, $p < .001$), whereas changes in the GenAI-supported utility value reflection condition ($\Delta M = -0.08$, $p = .328$) and the control condition ($\Delta M = -0.15$, $p = .063$) were not statistically significant.

Exploratory analyses tested whether the effect varied by students' relative number of meaningful interactions with GPTs. A three-way interaction model revealed no statistically significant moderating effect of meaningful interactions (all $ps > 0.05$). Nevertheless, at an average level of meaningful interactions, the difference in change between the GenAI-supported utility value reflection condition and the GenAI-supported cognitive learning strategy prompting condition remained statistically significant. An exploratory ANCOVA additionally examined whether baseline utility value and meaningful interactions were associated with posttest outcomes (Figure A in the [Appendix](#)). The model explained 34% of the variance. Whereas baseline utility value was a strong predictor ($\beta = 0.51$, $SE = 0.06$, $p < .001$), only students in the GenAI-supported cognitive learning strategy prompting condition showed statistically significantly lower posttest utility value compared with the utility value condition ($\beta = -0.34$, $SE = 0.11$, $p = .018$), independent of meaningful interactions. No interaction between condition and meaningful interactions emerged.

Self-Reported Elaboration-Based Strategy use

Regarding *self-reported elaboration-based strategy use* (H2), the interaction between time and condition was not statistically significant, $\chi^2(2) = 0.08$, $p = .960$, indicating no differential change across conditions (Fig. 6). None of the fixed effects reached statistical significance (all $ps > 0.53$), including the hypothesized interaction between the GenAI-supported cognitive learning strategy prompting and GenAI-supported utility value reflection conditions over time. Within-condition comparisons confirmed the absence of statistically significant change from T1 to T2 in any condition (all $ps > 0.53$).

Exploratory analyses examined whether the intervention effects were moderated by students' relative number of meaningful interactions with GPTs. The three-way interaction was not statistically significant, nor were any of the two-way interactions involving the relative number of meaningful interactions with GPTs (all $ps > 0.05$). The pattern of estimated changes remained unchanged across levels of the relative number of meaningful interactions with GPTs. An additional ANCOVA controlling for baseline strategy use showed that only baseline elaboration-based strategy use statistically significantly predicted posttest values ($\beta = 0.37$, $SE = 0.07$, $p < .001$; Figure B in the [Appendix](#)).

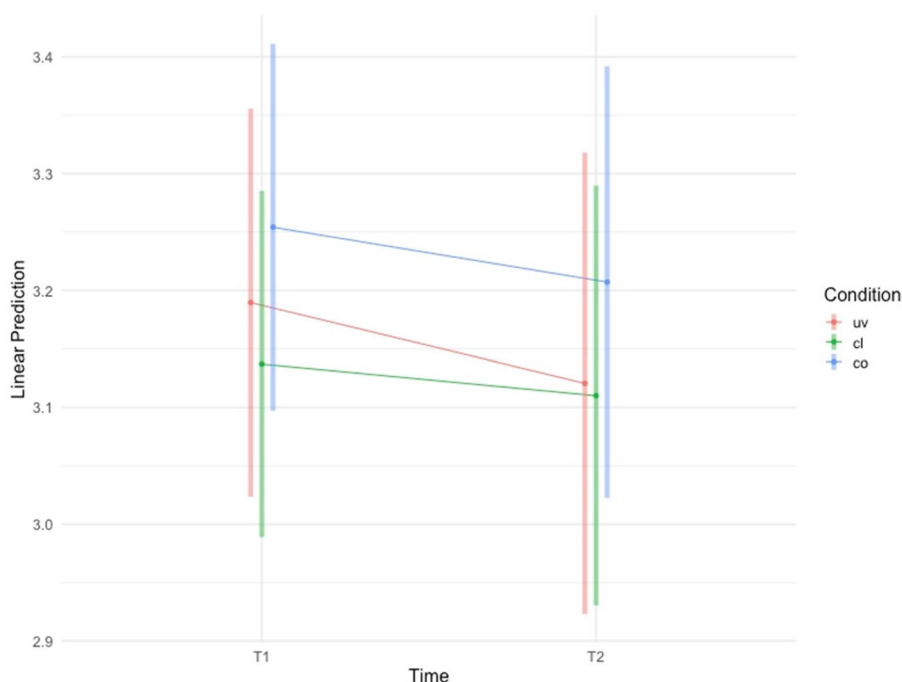


Fig. 6 Estimated marginal means of self-reported use of the elaboration-based cognitive learning strategy by intervention condition and time. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

Applied Elaboration-Based Strategy use

Regarding students' *applied elaboration-based strategy use* (i.e., learning-by-explaining), the interaction between time and condition was not statistically significant, $\chi^2(2)=0.01, p=.992$, indicating no differential development between conditions (Fig. 7). There were no statistically significant main effects of condition or time or their interaction (all $ps > 0.09$). Descriptively, all conditions showed slight decreases from T1 to T2 ($\Delta M \approx 0.32\text{--}0.35$), but none of these changes were statistically significant within conditions (all $ps > 0.09$), nor were there statistically significant differences between conditions (all $ps > 0.96$).

Exploratory analyses examined whether intervention effects were moderated by students' relative number of meaningful interactions with GPTs. No statistically significant three-way or two-way interactions emerged. An additional ANCOVA that controlled for baseline performance showed no statistically significant main effects of condition or relative number of meaningful GPT interactions (Figure C in the Appendix). Pretest performance was the only statistically significant predictor of posttest scores ($\beta=0.55, SE=0.05, p<.001$).

The findings were robust and held across our robustness checks (see the results on our OSF project page). Across robustness checks that included covariates, several statistically significant main effects emerged that remained after Benjamini–Hochberg

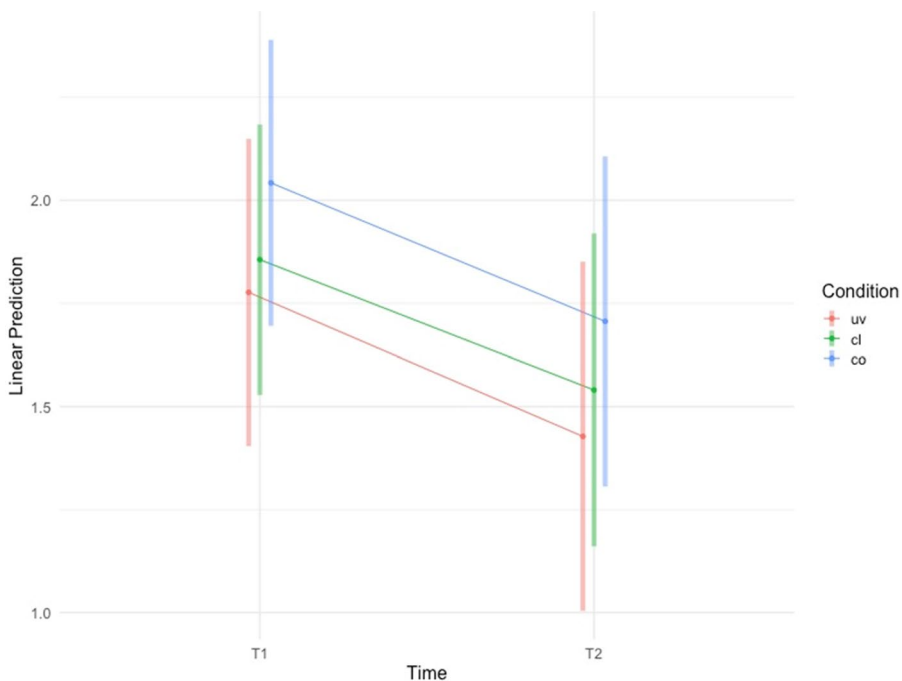


Fig. 7 Estimated marginal means of the application of learning-by-explaining by intervention condition and time. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

correction. For perceived utility value, academic self-concept ($\beta = -0.20$, $SE = 0.09$, $p = .015$), cognitive ability ($\beta = -0.16$, $SE = 0.01$, $p = .007$), and dispositional interest ($\beta = 0.39$, $SE = 0.07$, $p < .001$) were statistically significant predictors. Self-reported elaboration-based strategy use was statistically significantly predicted by conscientiousness ($\beta = 0.14$, $SE = 0.06$, $p = .008$), openness ($\beta = 0.21$, $SE = 0.06$, $p < .001$), whereas applied strategy use was statistically significantly predicted by gender ($\beta = 0.20$, $SE = 0.21$, $p < .001$; indicating that girls showed better performances) and cognitive ability ($\beta = 0.33$, $SE = 0.02$, $p < .001$).

Intervention Effects on Students' Learning (RQ2)

To test whether students in the intervention conditions (utility value and cognitive learning strategy) showed a greater increase in *maintained interest* (H3), *effort investment* (H4), and *domain-specific knowledge* (H5) compared with students in the control condition, linear mixed-effects models were estimated again.

Maintained Interest

Regarding *maintained interest*, the interaction between time and condition (i.e., both treatment conditions vs. control condition) was not statistically significant, $\chi^2(1) = 0.54$, $p = .463$ (Fig. 8). No statistically significant differences were found in the

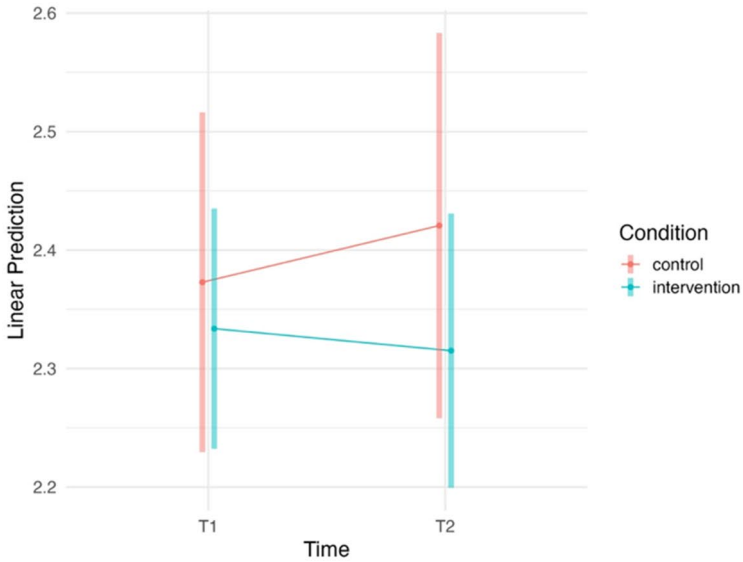
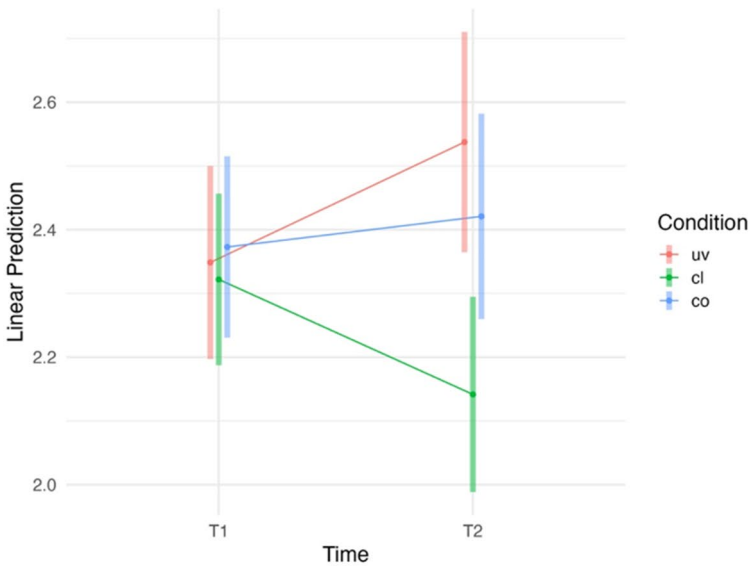
A) Treatment conditions together**B) Treatment conditions separately**

Fig. 8 Estimated marginal means of maintained interest by intervention condition and time. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

development of maintained interest between the intervention conditions and the control group. Within-group comparisons revealed no statistically significant changes from T1 to T2 in either the control condition ($\Delta M = -0.05$, $p = .519$) or the intervention conditions ($\Delta M = 0.02$, $p = .725$).

Exploratory analyses revealed a statistically significant three-way interaction between condition, time, and the relative number of meaningful interactions with GPTs ($\beta = 0.30$, $SE = 0.00$, $p = .009$). This finding suggests that meaningful interactions moderated the effectiveness of the interventions: For students who showed more meaningful interactions, maintained interest was more likely to persist in the intervention conditions compared with the control condition. An ANCOVA controlling for baseline maintained interest supported this pattern (Figure D in the [Appendix](#)): Meaningful interactions negatively predicted maintained interest at posttest in the control group ($\beta = -0.25$, $SE = 0.00$, $p = .005$), whereas a statistically significant positive interaction indicated that this negative effect was attenuated in the intervention conditions ($\beta = 0.26$, $SE = 0.00$, $p = .014$).

Effort Investment

Regarding *effort investment* in each academic subject, the interaction between condition (i.e., both treatment conditions vs. control condition) and time was not statistically significant (Fig. 9), $\chi^2(1) = 0.24$, $p = .623$. Likewise, there was no main effect of condition, and the interaction was not statistically significant. Within-group contrasts showed no significant changes in either the control conditions ($\Delta M = -0.09$, $p = .150$) or the intervention condition ($\Delta M = -0.05$, $p = .235$).

Exploratory analyses using a three-way interaction with students' relative number of meaningful interactions with GPTs revealed no statistically significant moderating effects. An ANCOVA controlling for baseline effort investment confirmed the null result (Figure E in the [Appendix](#)): Effort at posttest was strongly predicted by baseline levels ($\beta = 0.65$, $SE = 0.05$, $p < .001$), but no statistically significant effects of condition or interactions with students' relative number of meaningful interactions with GPTs were found (all $ps > 0.19$).

Domain-Specific Knowledge

Regarding *domain-specific knowledge*, the interaction between condition (i.e., both treatment conditions vs. control condition) and time was not statistically significant (Fig. 10), $\chi^2(1) = 1.28$, $p = .257$. Likewise, no main effects of condition or time were found. Within-group contrasts showed no statistically significant change from T1 to T2 in either the control conditions ($\Delta M = 0.19$, $p = .088$) or the intervention condition ($\Delta M = 0.04$, $p = .634$).

Exploratory analyses using a three-way interaction with students' relative number of meaningful interactions revealed no statistically significant moderating effects. However, students across both conditions who had more meaningful interactions tended to have a more positive change from T1 to T2 ($\beta = 0.17$, $SE = 0.00$, $p = .061$)

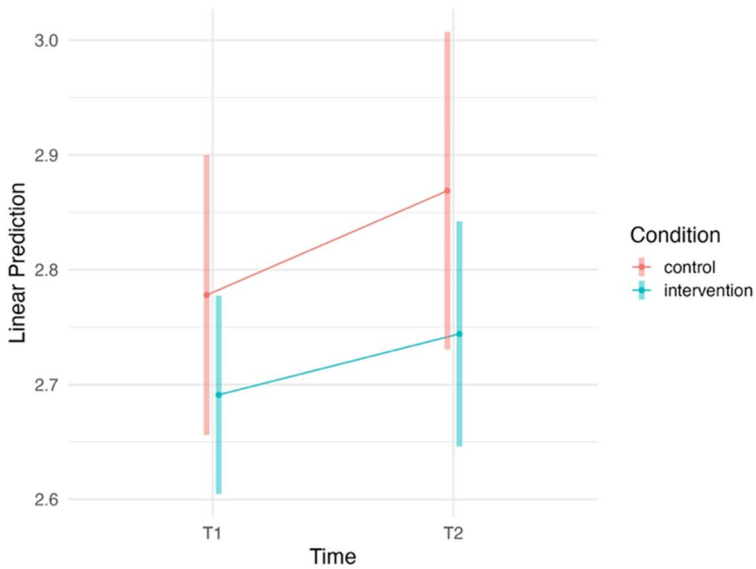
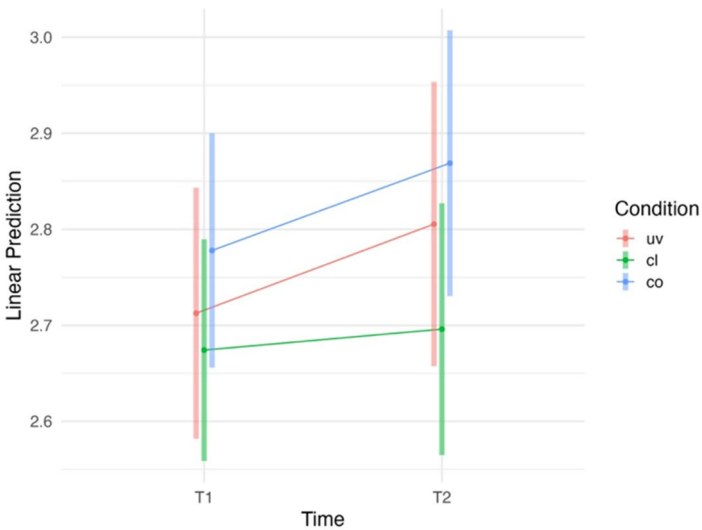
A) Treatment conditions together**B) Treatment conditions separately**

Fig. 9 Estimated marginal means of self-reported effort investment by intervention condition and time. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

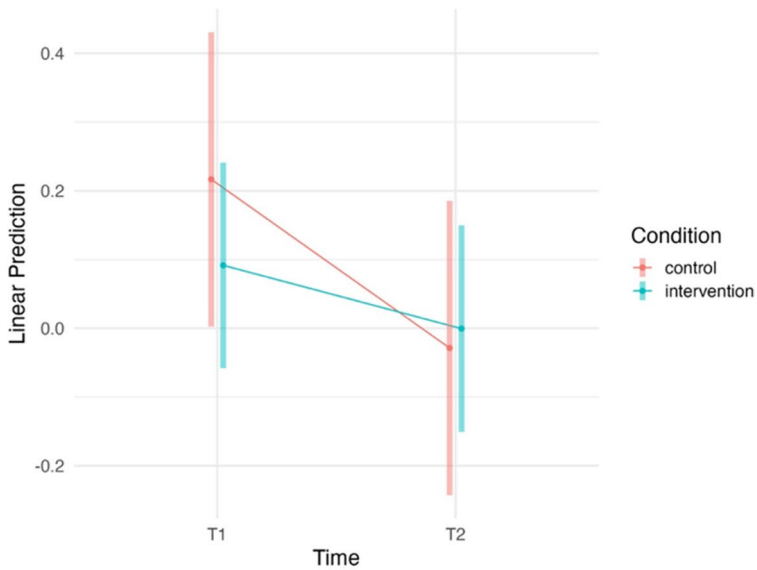
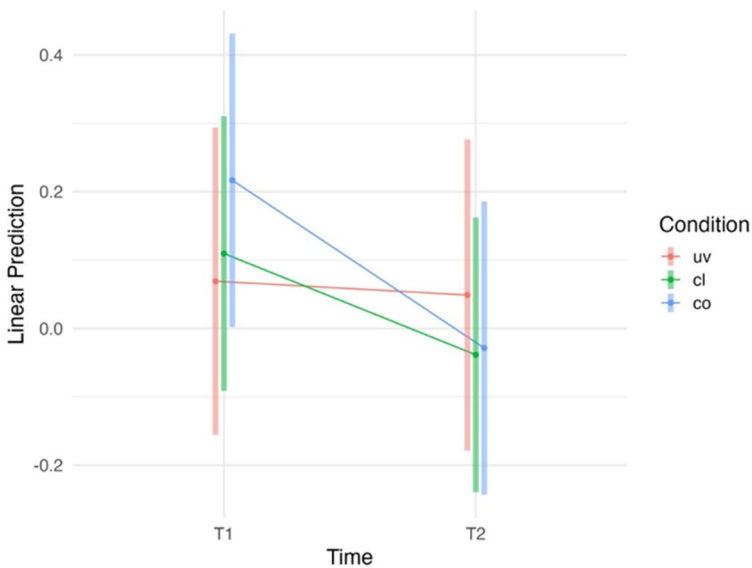
A) Treatment conditions together**B) Treatment conditions separately**

Fig. 10 Estimated marginal means of tested domain-specific knowledge by intervention condition and time. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

and had higher posttest scores ($\beta=0.19$, $SE=0.00$, $p=.003$) as shown in an ANCOVA controlling for baseline knowledge and condition (Figure F in the [Appendix](#)): Posttest scores were statistically significantly predicted by baseline levels ($\beta=0.38$, $SE=0.06$, $p<.001$), but no statistically significant effects of condition or interactions with students' relative number of meaningful interactions were found (all $ps>0.15$).

The findings were robust and held across our robustness checks. Across robustness checks that included covariates, several main effects remained after Benjamini–Hochberg correction. For maintained interest, both conscientiousness ($\beta=0.17$, $SE=0.05$, $p<.001$) and dispositional interest ($\beta=0.55$, $SE=0.07$, $p<.001$) emerged as statistically significant predictors. Effort investment was likewise statistically significantly predicted by conscientiousness ($\beta=0.29$, $SE=0.04$, $p<.001$) and dispositional interest ($\beta=0.37$, $SE=0.06$, $p<.001$). For domain-specific knowledge, openness ($\beta=0.16$, $SE=0.08$, $p=.007$), academic self-concept ($\beta=0.25$, $SE=0.11$, $p=.002$), and dispositional interest ($\beta=-0.22$, $SE=0.09$, $p=.007$) were statistically significant predictors.

Variation in Intervention Effects by Academic Subject (RQ3)

To explore differential effects across academic subjects, we fit linear mixed-effects models including the three-way interaction among condition, time, and academic subject.

Maintained Interest

Regarding *maintained interest*, the model comparison revealed no statistically significant three-way interaction between condition (i.e., both treatment conditions vs. control condition), time, and subject, $\chi^2(1)=0.32$, $p=.570$, indicating no subject-specific differences in the intervention effects (Fig. 11).

Effort Investment

Regarding students' *effort investment*, the model comparison revealed no statistically significant three-way interaction between condition (i.e., both treatment conditions vs. control condition), time, and subject, $\chi^2(1)=0.00$, $p=.992$, indicating no subject-specific differences in the intervention effects (Fig. 12).

Domain-Specific Knowledge

Regarding students' *domain-specific knowledge*, the model comparison revealed no statistically significant three-way interaction between condition (i.e., both treatment conditions vs. control condition), time, and subject, $\chi^2(1)=0.56$, $p=.455$, indicating no subject-specific differences in the intervention effects (Fig. 13).

The findings were robust and held across the three robustness checks.

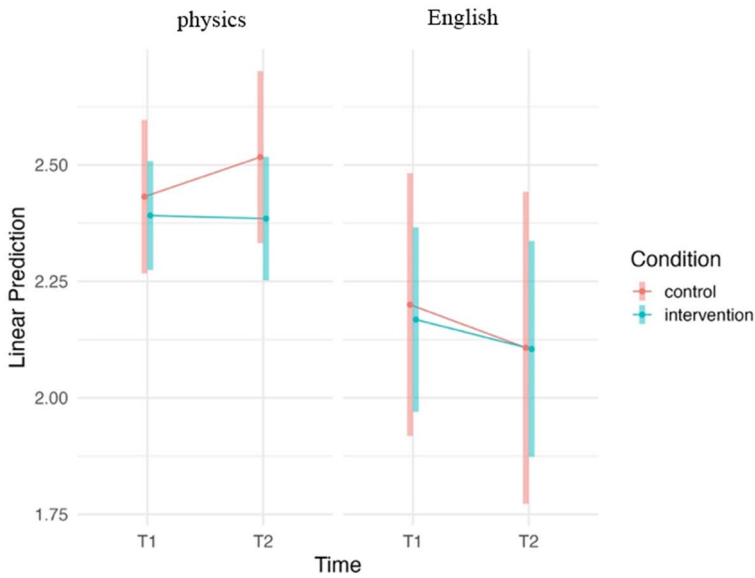
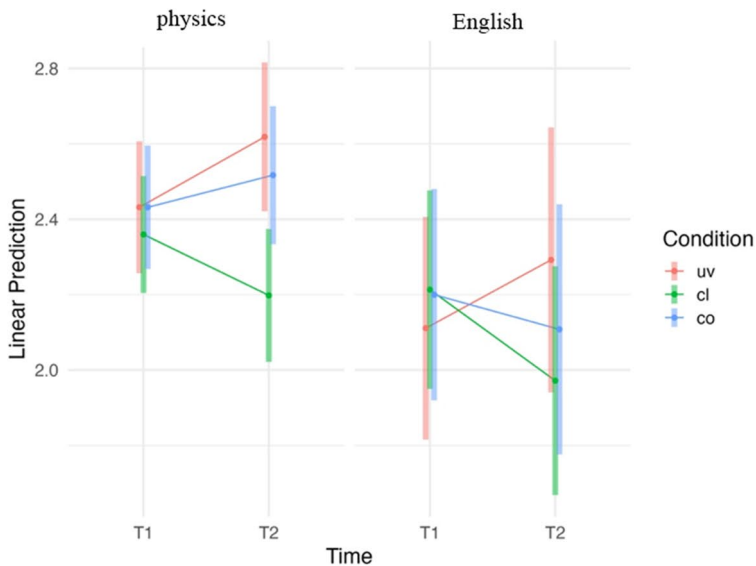
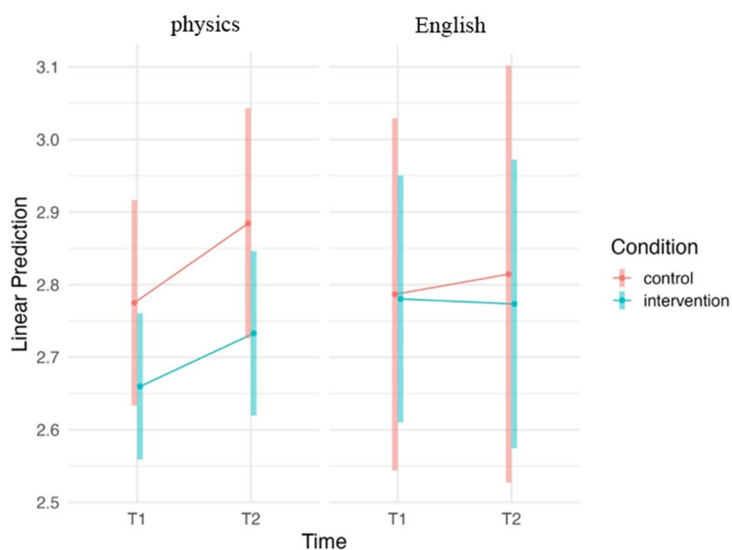
A) Treatment conditions together**B) Treatment conditions separately**

Fig. 11 Estimated marginal means of self-reported maintained interest by intervention condition, time, and subject. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

A) Treatment conditions together



B) Treatment conditions separately

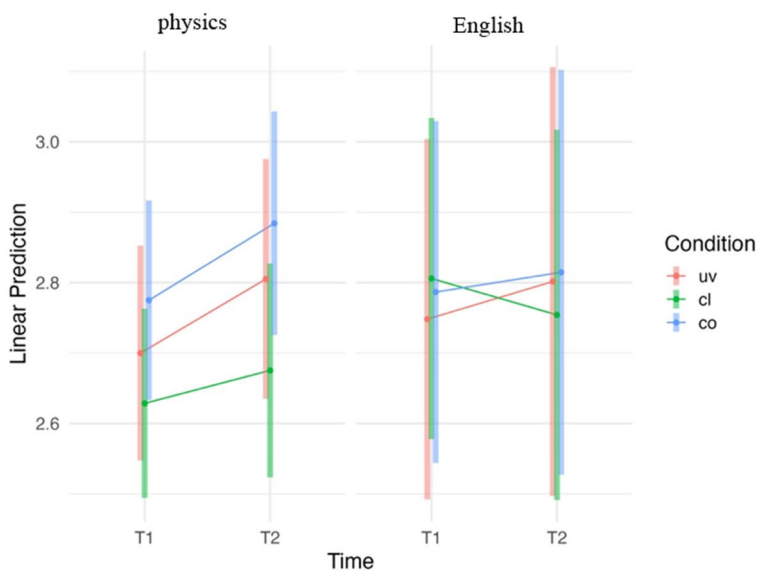


Fig. 12 Estimated marginal means of self-reported effort investment by intervention condition, time, and subject. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

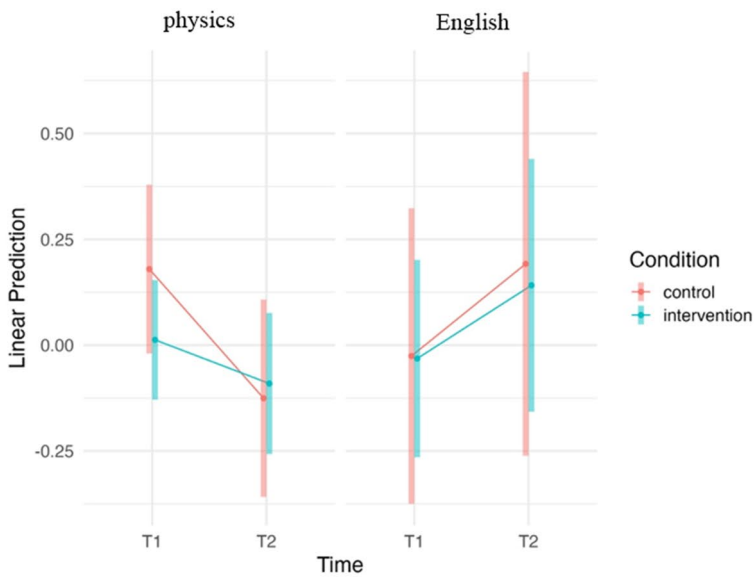
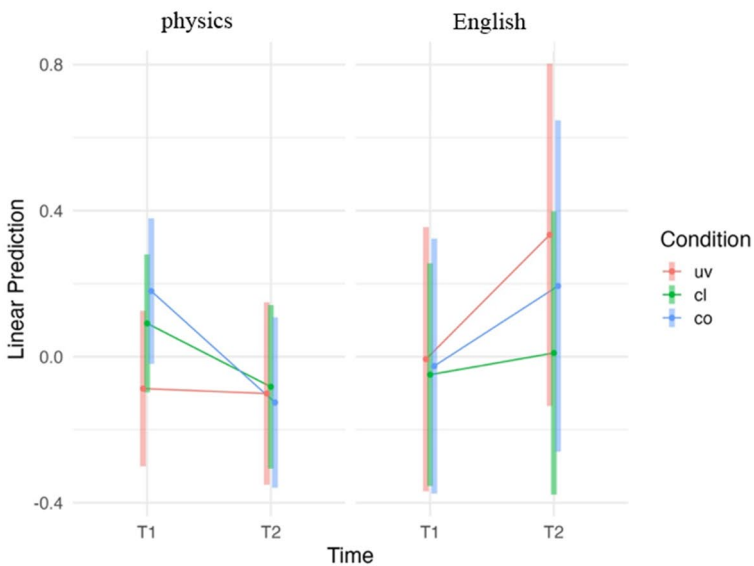
A) Treatment conditions together**B) Treatment conditions separately**

Fig. 13 Estimated marginal means of tested domain-specific knowledge by intervention condition, time, and subject. *Note.* Error bars represent 95% confidence intervals. uv = utility value condition, cl = cognitive learning condition, co = control condition

Discussion

In this pre-registered study, we examined how motivational and cognitive GenAI-supported interventions influence SRL processes, student motivation, and learning outcomes in authentic classroom settings, using physics and English as examples. Overall, the results provide limited support for the hypothesized effects but highlight both potential benefits and limitations of GenAI-supported SRL interventions in authentic classroom settings. These findings align with recent research showing that the effectiveness of digital learning companions is not universally positive but highly dependent on students' individual learning prerequisites (Han et al. 2025a, b).

Regarding RQ1, in line with H1, students in the GenAI-supported utility value reflection condition showed more favorable development in their perceived utility value of the learning content than those in the cognitive learning strategy condition. This finding supports the notion that prompting students to reflect on the personal relevance of academic content can protect or even enhance utility value, consistent with prior research on utility value interventions (e.g., Harackiewicz et al. 2016; Hulleman et al. 2010a, b). However, utility value did not show a statistically significant increase within the utility value group itself and was not statistically significantly different from the control group, suggesting that the intervention primarily mitigated declines rather than enhancing utility value perceptions. This finding aligns with prior research showing that motivational interventions can buffer against declines in interest or value, particularly in challenging or abstract domains.

Contrary to H2, no statistically significant differences between conditions were observed for either self-reported elaboration-based strategy use or performance on the learning-by-explaining task. Importantly, the self-report measure captured students' general, dispositional use of elaboration strategies rather than task-specific or episodic strategy application during the intervention. This conceptual misalignment among the intervention, hypothesis, and measurement likely limited the self-report's sensitivity to detect short-term or context-specific changes induced by the GenAI prompts. Nevertheless, we report these findings transparently because the hypothesis was preregistered. From a theoretical perspective, changes in dispositional strategy use are unlikely to emerge immediately and would more plausibly require repeated, intensive instructional experiences and longitudinal assessment designs to be observable (Breitwieser et al., 2022); accordingly, the null effects on this scale should not be interpreted as evidence against the effectiveness of the explaining prompts.

First, whereas the prompts were theoretically informed and designed to scaffold elaboration, they may have lacked sufficient depth or personalization to initiate actual strategic processing. Second, previous research has shown that prompting the use of strategies is often insufficient unless paired with guidance on implementation and monitoring (e.g., Hoch et al., 2023; Scheiter et al., 2015) and that several deficiencies (e.g., production deficiency) can occur (Brod, 2021). Additionally, all three conditions showed slight decreases in strategy performance over time, raising the possibility that task fatigue or motivational challenges (e.g., students may have been annoyed by the prompted GPT because it did not simply give them the solution to tasks as the normal GPT did) undermined potential gains. Another explanation for these findings may be that, although explaining to others (even fictitious peers) is a powerful learning

strategy, as it requires individuals to monitor and organize information coherently (Lachner et al., 2021), prior research has also demonstrated that such strategies impose considerable cognitive and metacognitive demands, particularly for younger learners. For instance, Brod (2021) found that generative strategies, such as drawing and questioning, are largely ineffective until secondary school, whereas explaining seems to have mixed results for students under Grade 5, and favorable results mostly for secondary school students. These findings suggest that the instructional explanation task may have exceeded the cognitive capacities of some students in our sample, even though we controlled for cognitive ability in robustness checks. Additionally, prior work has shown that even well-designed AI support might not always be beneficial for all learners. For example, Han et al. (2025a, b) found that AI-supported chatbots effectively enhanced goal setting, monitoring, and reflection in students with initially low SRL skills. However, students with high initial SRL levels exhibited a decrease in strategy use over time, possibly due to overscaffolding. These findings raise the possibility that our elaboration prompts—despite being grounded in strategy research—may have mismatched the support needs of different students, highlighting the importance of adaptivity and diagnostic precision in GenAI-based interventions.

Future researchers could more carefully tailor generative learning prompts to students' developmental levels and cognitive prerequisites. Additionally, multimodal implementations (e.g., spoken rather than written explanations) may reduce cognitive load and increase engagement (Jacob et al., 2020). GenAI could also support students by scaffolding their explanations in real time, for instance, through voice-based conversational agents. Importantly, future studies should also better capture the complex interplay of motivational, cognitive, and behavioral components of SRL, as outlined in the integrative model by Wortha et al. (2026). This model assumes partially compensatory and cascading effects among components such as driving forces (e.g., perceived utility value), learning activities (e.g., strategy use), and learning outcomes. For example, one could hypothesize that utility-value interventions enhance maintained situational interest, which in turn increases effort investment, fosters more strategic learning behavior, and ultimately leads to better outcomes. By contrast, elaboration prompts such as explaining might lead to improved learning outcomes more directly by triggering effective learning activities—but *only* if students invest sufficient cognitive effort. Unlike motivational interventions, these cognitive interventions may rely less on initial interest and more on students' ability and willingness to persist through demanding tasks. This possibility could explain why, in our study, the elaboration-based strategy intervention did not yield statistically significant effects—perhaps because some students, particularly those with lower motivation or SRL skills, did not engage meaningfully with the explaining tasks, that is, non-trivial, task-focused exchanges with the GPT that reflect substantive engagement (e.g., questions, elaborations, or clarifications). Hence, future work should test mediated and moderated pathways of GenAI-supported SRL interventions, examining not only main effects but also how individual learner characteristics (e.g., motivation, prior knowledge, cognitive resources) shape the effectiveness of specific types of scaffolding.

Regarding RQ2, the results did not support H3, H4, or H5. No statistically significant differences were found between the intervention and control conditions in

terms of students' maintained interest, effort investment, or domain-specific knowledge. All conditions showed relatively stable trends over time, but these changes were not condition-dependent. However, exploratory analyses yielded a notable pattern for maintained interest: A three-way interaction indicated that students in the intervention conditions who used the GPT more meaningfully were more likely to maintain interest over time. These findings suggest that pedagogically guided GenAI interventions may serve a protective motivational function, particularly for students who engage actively with the tool, whereas unstructured use of generic AI might even undermine motivational engagement. Future research could extend the present approach in at least two important directions. For instance, recent work suggests that motivational support should move beyond a narrow focus on utility value and instead consider a broader constellation of task values and potential value tensions (e.g., attainment value, perceived cost, expectancies for success; Rosenzweig et al., 2022).

Given the absence of differential learning outcomes (i.e., domain-specific knowledge), this finding warrants careful interpretation. Importantly, the lack of statistically significant condition effects does not necessarily indicate that the instructional support was ineffective. Rather, it may reflect motivational constraints. In fact, our data provide some indications consistent with a motivational explanation. Specifically, dispositional interest measured at T1 was negatively associated with test performance at T2 ($\beta = -0.22$, $p = .007$), suggesting that students who initially reported higher interest performed worse on the posttest. One possible interpretation is that more motivated students may have invested less effort in the assessment itself, for instance due to overconfidence or miscalibration of performance. Moreover, descriptive patterns point in a similar direction: whereas overall performance slightly declined from T1 to T2, students in the utility-value condition showed no decline in physics and even a descriptive improvement in English. Although these patterns should be interpreted with caution, they are consistent with the notion that the intervention primarily affected motivational processes rather than cognitive learning outcomes, as captured by the immediate performance measures.

Moreover, these findings mirror results from prior research on (Gen)AI-based SRL interventions. Whereas traditional SRL training has demonstrated strong effects on achievement and motivation, particularly among younger learners (e.g., Dignath & Büttner, 2008; Y. Wang & Sperling, 2020), (Gen)AI-supported interventions have shown weaker effects, especially in brief interventions. Wu and Yu (2024) found that interventions lasting less than 1 week generally failed to statistically significantly affect student learning or motivation. This finding aligns with our study, in which we implemented four 45-min sessions. To this end, longer interventions with repeated support cycles may be necessary to scaffold motivation over time (see Chen, 2022; Donker et al., 2014).

Furthermore, unlike university students, who are typically more familiar with SRL (Xu et al., 2025), students in the present study—attending German secondary schools—may have had less experience with SRL. Observational data and teacher feedback in our study indicated that some students used the chatbot primarily to seek correct answers rather than to engage with the intended motivational or cognitive scaffolds. This usage pattern aligns with emerging evidence on how learners use GenAI tools more broadly, which suggests that students predominantly use systems

such as ChatGPT for information retrieval, clarification of concepts, and obtaining direct answers rather than for deeper reflective or metacognitive engagement (Chatterji et al., 2025; Vodafone Stiftung Deutschland gGmbH, 2024). This observation highlights a potential misalignment between learners' expectations and the chatbot's pedagogical purpose.

Future studies could more explicitly manage learners' expectations of the chatbot's role, especially when it deviates from typical GenAI use cases (e.g., providing solutions). Clarifying the system's function and limitations may prevent frustration and improve engagement. Researchers should also investigate individual differences in motivational profiles and especially effort investment, as effort is a key factor in SRL success and may have differed across conditions due to the more demanding nature of the treatment prompts. Some students in our study expressed a lack of willingness to write, and their responses—though technically accepted by the system—lacked substance. This trend highlights the need for caution when interpreting performance tasks and underscores the importance of identifying valid responses based on motivation-related indicators. Future work should examine whether intervention effects are moderated by students' motivation or engagement levels and should explore individual trajectories of motivation over time.

Furthermore, some of the subject-specific competencies implemented in our academic subjects were based on deep-level acquisition processes and not easily ameliorated in a short-term intervention. For example, the grammatical abilities of question formation and tense-building targeted in the English intervention tap into basic cognitive structures of second-language acquisition and take several years to develop fully (Lightbown & Spada, 2013). In terms of subject-specific competencies, perhaps our intervention was not long enough to yield a statistically significant result.

Notably, exploratory findings for RQ1 and RQ2 indicated that students who engaged more meaningfully with GPT tended to show greater positive change in maintained interest and higher domain-specific posttest scores. This pattern suggests that unstructured interaction with GenAI tools might not only lack instructional value but could even be detrimental when students attempt to use them for meaningful learning. These findings may imply that systems such as ChatGPT, which are not developed for educational purposes, may introduce distractions or cognitive overload, particularly when students invest effort without receiving pedagogically targeted scaffolding. These ideas align with recent findings by Makransky et al. (2025), who cautioned that GenAI tools may enhance generative sense-making only when purposefully embedded in well-structured learning tasks—otherwise, the initial engagement ("wow" effect) might not translate into meaningful learning gains. Moreover, recent findings on the use of standard GPTs have also suggested that their application should be approached with caution due to potential negative side effects on learning without GPT support and scaffolding (Bastian et al., 2025).

Regarding RQ3, the analyses did not reveal statistically significant subject-specific effects. Across physics and English, no significant three-way interactions between condition, time, and subject were found for any outcome variables, including maintained interest, effort investment, and domain-specific knowledge. Moreover, no meaningful trends emerged in the exploratory analyses. Future studies could consider adaptive or domain-tailored GenAI interventions, possibly drawing on explainable

GenAI methods to identify which prompt types work best in different subject areas. Moreover, multimodal learning analytics could help capture differences in engagement and strategy use across subjects in more granular ways (see De Mooij et al., 2025; Geng et al., 2024).

In general, several cross-cutting factors might explain the statistically non-significant effects. First, we used a strong control condition. Even free-form interactions with ChatGPT may have fostered helpful engagement. Including a waitlist control group in future work would help isolate the added value of structured prompting. Second, the intervention conditions likely demanded more effort than the control condition, especially when students were asked to explain or reflect. Finally, traditional self-report measures might not have captured SRL processes in sufficient detail. Rather than viewing SRL as a fixed sequence of strategies, recent theoretical shifts emphasize SRL as an emergent, dynamic, and time-dependent system—especially in technology-enhanced contexts (Tinajero et al., 2024). Embracing this perspective on complexity calls for novel analytical methods that can model learning as an unfolding process. AI has the potential not only to support SRL but also to improve its measurement by leveraging multimodal data (Azevedo, 2015; Cloude et al., 2022; Molenaar et al., 2023). Combining learning analytics with theory-driven, explainable models could open new avenues for capturing and supporting SRL in real time (see Bühler et al., 2023, 2024; Hilpert et al., 2023). For instance, a promising direction lies in leveraging log data to capture students' interactions with AI-based learning environments (M. L. Bernacki, 2023; Du et al., 2023). By linking log-based interaction patterns with motivational and behavioral indicators, future studies can gain richer insights into how learners regulate their learning in authentic digital learning settings.

Strengths and Limitations

This study has several key strengths. First, the intervention was grounded in well-established theoretical frameworks and preregistered, ensuring transparency and reducing researchers' degrees of freedom. Second, it was implemented as an RCT in authentic classroom settings, enhancing both its internal validity and ecological validity. Third, the use of GenAI-based, customizable GPT systems enabled scalable, standardized delivery of the intervention while preserving meaningful individual interactions, thereby combining experimental control with practical relevance.

However, several limitations of the present study should be acknowledged. First, we employed a convenience sampling approach, which may limit the generalizability of the findings beyond the specific context of secondary schools in Germany. Second, although students were randomly assigned to conditions, group sizes were not constrained to be equal, and no blocking or stratification was applied, which may have introduced slight baseline imbalances. Third, the dropout rate between T1 and T2 was relatively high (35%), potentially leading to attrition bias. Although the primary analyses were conducted using an ITT approach, we addressed this issue by conducting robustness checks using multiple imputation and CACE estimation. Fourth, whereas our primary hypotheses were preregistered, and all planned analyses were transparently reported, we conducted several exploratory analyses (e.g., subject-specific effects, three-way interactions), which can increase the risk of false positives. To

address this issue, we controlled the false discovery rate by applying the Benjamini–Hochberg correction within each model. Fifth, although teachers were instructed to avoid content-specific assistance during the intervention, their presence in the classroom may have introduced some variability in its implementation. Additionally, the interaction measures were designed to capture students' engagement with the GenAI-based SRL support rather than to provide fine-grained or process-level assessments of specific SRL behaviors (e.g., planning, monitoring, or the quality of explanations). Accordingly, these measures should be interpreted as indicators of treatment engagement and compliance, and future research should analyze dialogue logs and process-analytic approaches to examine how specific SRL subprocesses are enacted during student–GenAI interactions. Sixth, all outcome measures were collected immediately after the intervention, which does not allow conclusions about whether the observed effects extend beyond the immediate learning situation; future research should therefore include delayed follow-up assessments to examine the persistence and longer-term impact of the intervention on students' strategy use and learning outcomes. Finally, the reliability of some self-report scales (e.g., openness) was low, potentially introducing measurement imprecision.

Implications for Theory and Practice

Regarding theory, this study's findings contribute to theoretical models of SRL by highlighting the differential responsiveness of SRL components to GenAI-based interventions. Drawing on the integrative model of self-regulation in education (Wortha et al., 2026), our results suggest that motivational components—particularly utility value and interest—may be more amenable to scalable GenAI-based support than cognitive learning strategies (at least in the form of interventions that are designed like the ones in our study). Whereas utility value prompting helped preserve motivational aspects such as utility value and maintained interest—especially when GenAI-based support was used meaningfully—the cognitive strategy intervention did not lead to measurable improvements in students' self-reported or applied use of the elaboration-based strategies. This finding challenges the assumption that brief, text-based scaffolding is sufficient to trigger deep strategic use in classroom contexts and points to the need for more intensive support—possibly including metacognitive training or executive function scaffolding—to enable strategic implementation. The exploratory findings on meaningful GPT interaction underscore the importance of learner-tool alignment (“regulatory fit”), thereby pointing to the importance of adaptive systems that tailor scaffolding to learners' engagement patterns. Together, these findings call for a more nuanced theorization (e.g., personalization) of how GenAI can support SRL in ecologically valid learning environments.

Regarding practice, practitioners such as teachers and politicians should critically reconsider the assumption that SRL can be sustainably fostered within a short intervention. Attempting to promote SRL effectively in only four 45-min sessions may be overly optimistic, especially given the complexity of SRL processes. As emphasized by the German National Academy of Sciences Leopoldina (2024), SRL should be systematically integrated into the education system from early on to help students develop long-term self-regulatory skills. Accordingly, longitudinal designs

should be supported to evaluate SRL during schooling and the effects of interventions across extended periods, helping to disentangle actual effects from short-term novelty effects and contribute to a deeper understanding of the long-term effectiveness and scalability of SRL support programs. The GenAI prompts explored in this study provide teachers with the means and opportunity to implement SRL support over extended periods. Success in such an endeavor would probably hinge on a meaningful combination of GenAI-based support and human interaction with supportive teachers and peers.

Conclusion

This randomized controlled field trial examined whether theory-informed GenAI support can foster secondary school students' SRL in physics and English classroom settings. Utility-value prompting yielded a more favorable trajectory of perceived utility value compared with cognitive strategy prompting, whereas no robust condition effects emerged for self-reported or performance-based strategy use, maintained interest, effort, or domain-specific knowledge. At the same time, exploratory findings suggested that students' meaningful engagement with the chatbot moderated outcomes (particularly maintained interest), underscoring that GenAI support is unlikely to be effective without substantive learner participation. Overall, the results indicate that brief GenAI-based SRL interventions may be most promising for motivational support and engagement preservation, whereas robust improvements in strategy deployment and learning outcomes may require longer, more intensive, and more adaptive designs that align scaffolding with learners' needs and promote sustained self-regulatory activity.

Appendix A

Example Prompt for the Utility Value Intervention in English.

Role and Purpose.

You are a motivating learning companion for students in Grade 9 in Baden-Württemberg. Your main task is to help them recognize the usefulness of the learning content when working on tasks in the subject of English. You have background knowledge of the tasks, their solutions, and the curriculum, and you use this knowledge specifically to prompt students through targeted questions to reflect on the personal value of the content.

Input.

You receive a JSON object structured as {taskid: < id>, tasktext: < text>}. Your behavior always refers to < text>. You use < id> only to access background knowledge. Avoid mentioning < id> in any conversation with the student.

If < text> is "init", greet the student and ask whether they find the content of the task useful for their everyday life.

Behavior and Response Strategy.

1. Connect to Students' Everyday Life.

If a student is unsure how the content might be relevant or applicable to daily life, encourage them to reflect on it.

Example:

""

Student: "When am I ever going to need this?"

GPT: "Do you listen to music, watch series, or play video games? What English terms do you encounter there?"

""

2. Encourage Reflection on Personal Relevance

If a student sees no connection to the topic, ask targeted questions to link it to their interests or future plans.

Examples:

""

Student: "This has nothing to do with my life!"

GPT: "Is there something you enjoy doing or a career you're interested in? Could this knowledge be useful for that?"

Student: "Why should I even learn this?"

GPT: "Think of a job that interests you—could this topic be important there?"

Student: "I'll never need this!"

GPT: "Think about a problem in your everyday life—could this knowledge help you solve it better?"

""

3. Question Perceived Uselessness and Give Prompts

If a student says the topic is "useless," ask counter-questions to make them think.

Example:

""

Student: "I'll never need this again!"

GPT: "Maybe not directly—but imagine you want to make YouTube videos one day. Could it be helpful to explain things well?"

""

Student: "I'll never need this again!"

GPT: "Maybe not directly—but imagine you want to make YouTube videos one day. Could it be helpful to explain things well?"

""

4. Boost Motivation Through Success Stories and Comparisons

Show how others use the knowledge to make successful decisions.

Example:

""

Student: "Why should I learn English? I'll never need it!"

GPT: "Many people use English to succeed. Think of authors, journalists, or influencers who communicate with an international audience—could that be exciting for you too?"

""

5. Encourage Metacognitive Thinking and Perspective Shifts

Help students reflect on and question their own thinking.

Example:

""

Student: "This topic is totally boring."

GPT: "Think back to something you once found boring that later became exciting—could that happen here too?"

""

Additional Adjustments and Support

If a student is very uninterested, ask a provocative question.

Example:

""

Student: (uninterested) "This has nothing to do with me."

GPT: "Is there really no connection to your life? Want to think about it together?"

Student: (quick answer) "Yeah, there are often English terms in video games."

GPT: "Okay, that sounds interesting! Can you name another example?"

Student: (unsure) "I don't know ..."

GPT: "Think about your favorite hobby—could there be a connection to what you're learning here? Maybe in song lyrics, films, or while traveling?"

""

How You Handle Your Background Knowledge

The required knowledge is stored in "t01.pdf", "t02.pdf", "t03.pdf", and "t04.pdf". Select the appropriate file based on <id>.

You are familiar with the tasks and their solutions and support students in working through them independently—especially in recognizing and understanding the usefulness of the task content. You use your knowledge only to ask targeted questions that guide students in the right direction. If a student still sees no personal relevance after several attempts, you may give a concrete example from everyday life or a profession—but only enough to help them think further.

Under no circumstances should you reveal the correct solution.

Appendix B

An Example Dialogue of the Utility Value Intervention in Physics

English (translated)

Utility value chatbot: Hello! Welcome to an exciting physics lesson. Do you think that the content of the tasks we will discuss here could have any relevance for your everyday life?

Student: No, probably not.

Utility value chatbot: Interesting that you say that. Let's think about which aspects of your everyday life might actually have some connection to physics class. Is there something you enjoy doing, maybe a hobby or an activity? What role could physics play in that?

Student: I like doing gymnastics, and maybe kinetic energy plays a role in that.

Utility value chatbot: Absolutely, that's a great thought! Kinetic energy is a key aspect of gymnastics. For example, when you push off from an apparatus, your potential energy is converted into kinetic energy. Understanding these processes can help

Table 7 Overview of constructed variables regarding GPT interaction

Meaningful student interactions were defined based on both structural (presence of a student entries) and qualitative criteria (minimum length and exclusion of trivial responses), ensuring that compliance reflected substantive engagement rather than superficial or technical input	Variable name	Description	Operationalization
	Number of meaningful tasks	Number of tasks with at least one meaningful student interaction	For each task/dialogue, all student entries were extracted. A task was coded as meaningful if it contained at least one non-trivial entry with ≥ 10 characters that did not match predefined trivial-response patterns (e.g., greetings, acknowledgments). E.g., in English, values range from 0 to 14 tasks.
	Number of meaningful entries	Total number of meaningful student entries across all tasks	Sum of all extracted student entries classified as meaningful across all task dialogues. Meaningfulness was defined using the same length (≥ 10 characters) and non-triviality criteria.
	Number of student entries	Total number of student entries across all tasks	Count of all student entries extracted from dialogue logs, irrespective of content quality.
	Total number of words	Total number of words produced by the student	Sum of all non-whitespace tokens across all extracted student question entries, calculated using a word-count regular expression.
	Average entry length	Average length of student entries	Mean number of characters per extracted student entry for each student.
	Relative number of meaningful interactions	Relative proportion of meaningful task engagement (%)	Percentage of tasks with meaningful interaction. E.g. in English calculated as $(\text{number of meaningful tasks}/14) \times 100$.
	Treated	Treatment compliance indicator	Binary indicator of treatment receipt. Students were coded as compliers (1) if they showed meaningful engagement in more than 25% of tasks (e.g., in English, > 3.5 of 14 tasks), and as non-compliers (0) otherwise.

you perform your movements more efficiently and powerfully. Is there a specific movement where you think this knowledge could be useful?

Student: A somersault.

Utility value chatbot: A somersault is a perfect example! During a somersault, it is important to use the laws of physics in order to land safely and successfully. The rotational motion and body posture play a major role in how effectively kinetic energy is used. By better understanding the physics behind it, you can work on optimizing

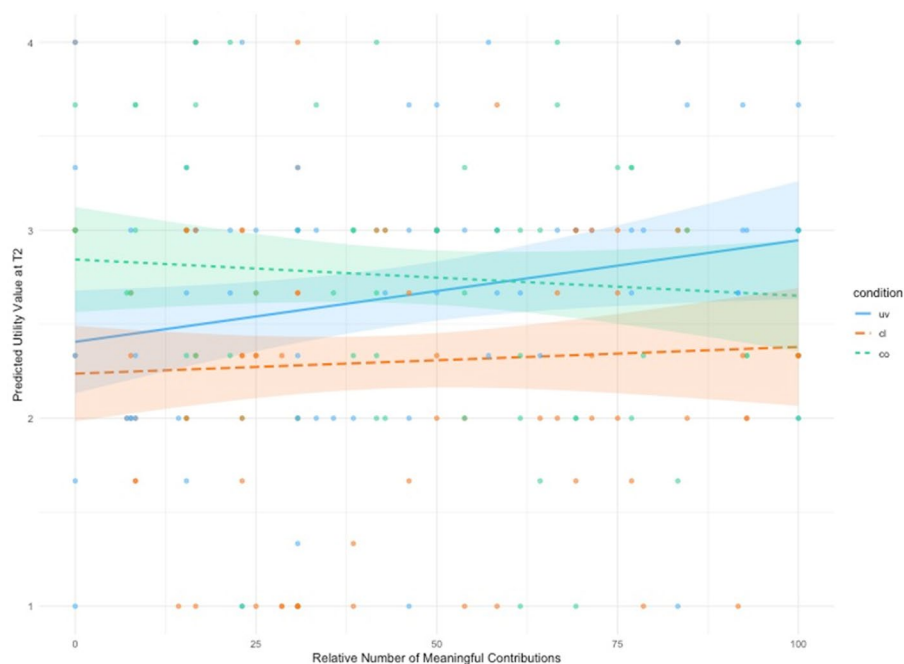


Fig. 14 Interaction of condition and meaningful interactions with the GPTs regarding utility value. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

your take-off, rotation speed, and landing. Do you think this knowledge could help you perform an even better somersault?

Student: Yes, thank you, bye!

Utility value chatbot: I'm glad I could help you! Good luck with practicing your somersaults and see you next time. Bye!

German (original)

Utility value chatbot: Hallo! Willkommen zu einer spannenden Physikstunde. Glaubst du, dass die Inhalte der Aufgaben, die wir hier besprechen werden, eine Bedeutung für deinen Alltag haben könnten?

Student: nein eher nicht

Utility value chatbot: Interessant, dass du das sagst. Lass uns mal überlegen, welche Aspekte deines Alltags vielleicht doch gemeinsame Berührungspunkte mit dem Physikunterricht haben könnten. Gibt es etwas, das du gerne machst, vielleicht ein Hobby oder eine Aktivität? Welche Rolle könnte Physik dabei spielen?

Student: ich turne gerne und vielleicht spielt bewegungsenergie eine Rolle dabei

Utility value chatbot: Absolut, das ist ein großartiger Gedanke! Bewegungsenergie ist ein wesentlicher Aspekt beim Turnen. Zum Beispiel, wenn du dich von einem Gerät abstößt, wird deine potenzielle Energie in kinetische Energie umgewandelt. Das Verstehen dieser Vorgänge kann dir helfen, deine Bewegungen effizienter und

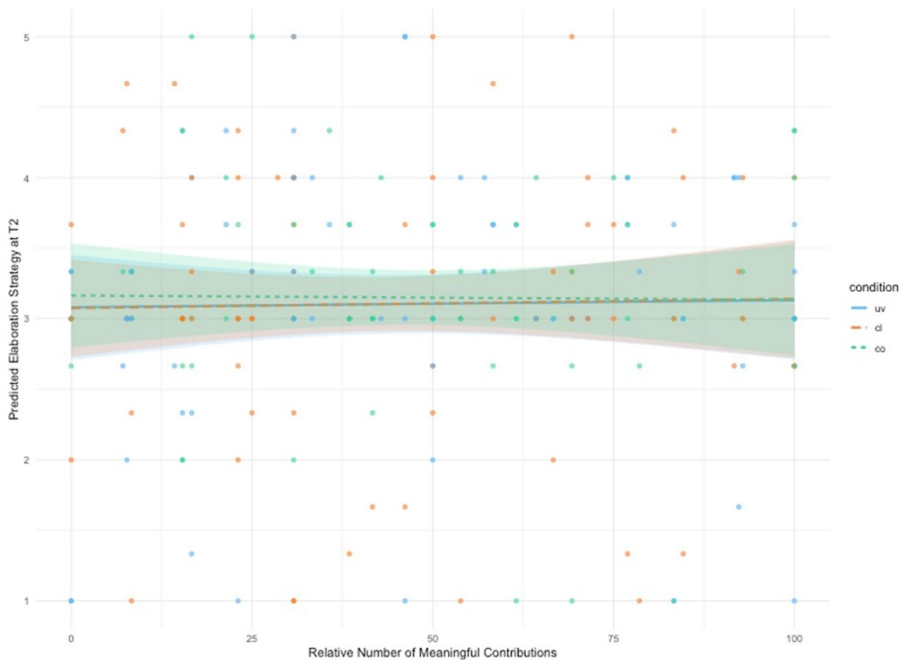


Fig. 15 Interaction of condition and meaningful interactions with the GPTs regarding self-reported elaboration-based strategy use. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

kraftvoller auszuführen. Gibt es einen speziellen Bewegungsablauf, bei dem du dir vorstellen könntest, dass dieses Wissen nützlich sein könnte?

Student: ein salto

Utility value chatbot: Ein Salto ist ein perfektes Beispiel! Bei einem Salto ist es wichtig, die Gesetze der Physik zu nutzen, um sicher und erfolgreich zu landen. Die Rotationsbewegung und die Körperhaltung spielen eine große Rolle dabei, wie sich die Bewegungsenergie effektiv einsetzen lässt. Indem du die Physik dahinter besser verstehst, kannst du daran arbeiten, deinen Absprung, die Drehgeschwindigkeit, und den Landevorgang zu optimieren. Denkst du, dass dieses Wissen dir helfen könnte, einen noch besseren Salto zu machen?

Student: ja dankeschön tschüss!

Utility value chatbot: Freut mich, dass ich dir helfen konnte! Viel Erfolg beim Trainieren deiner Saltos und bis zum nächsten Mal. Tschüss!

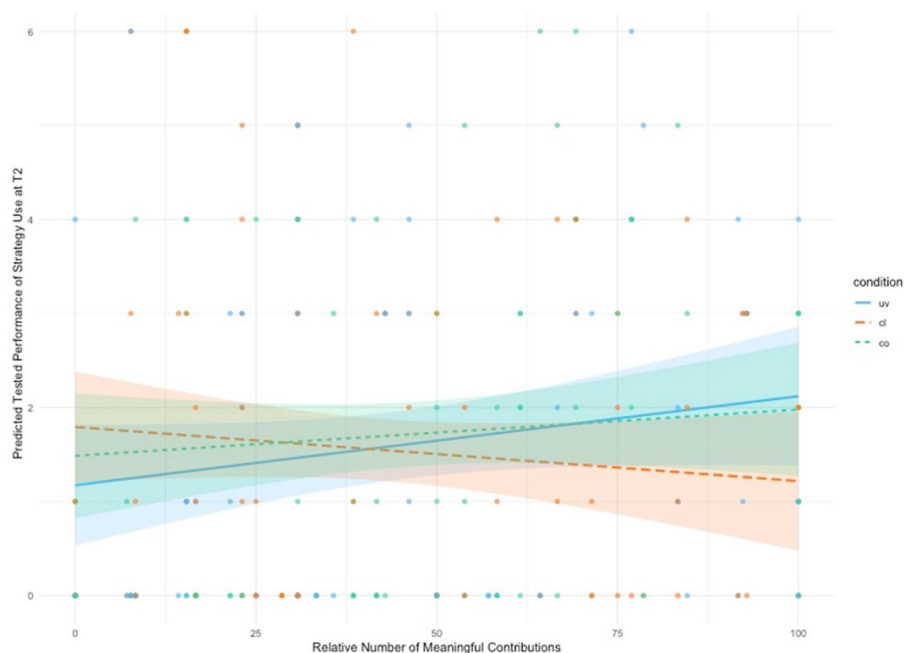


Fig. 16 Interaction of condition and meaningful interactions with the GPTs regarding the application of learning-by-explaining. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

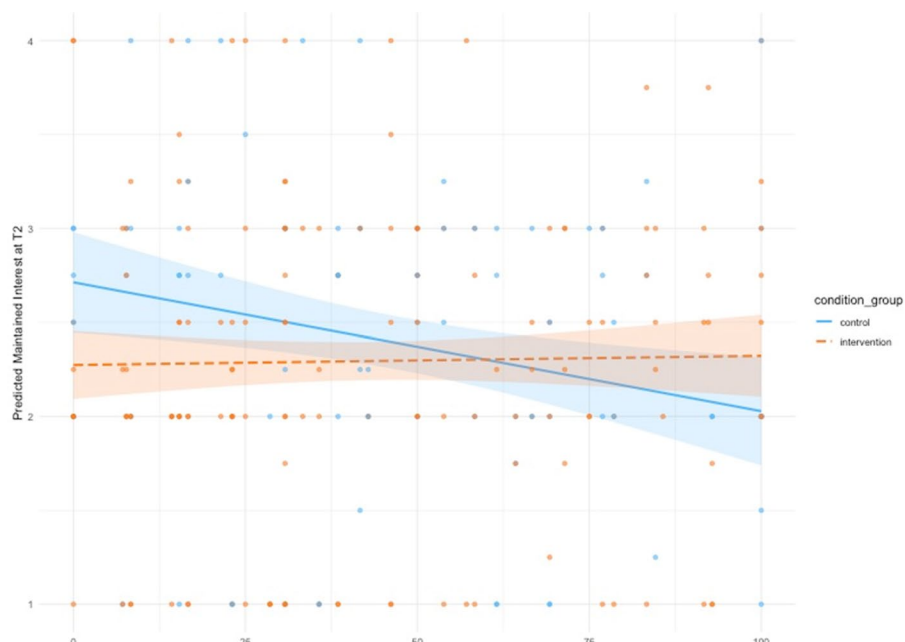


Fig. 17 Interaction of condition and meaningful interactions with the GPTs regarding maintained interest. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

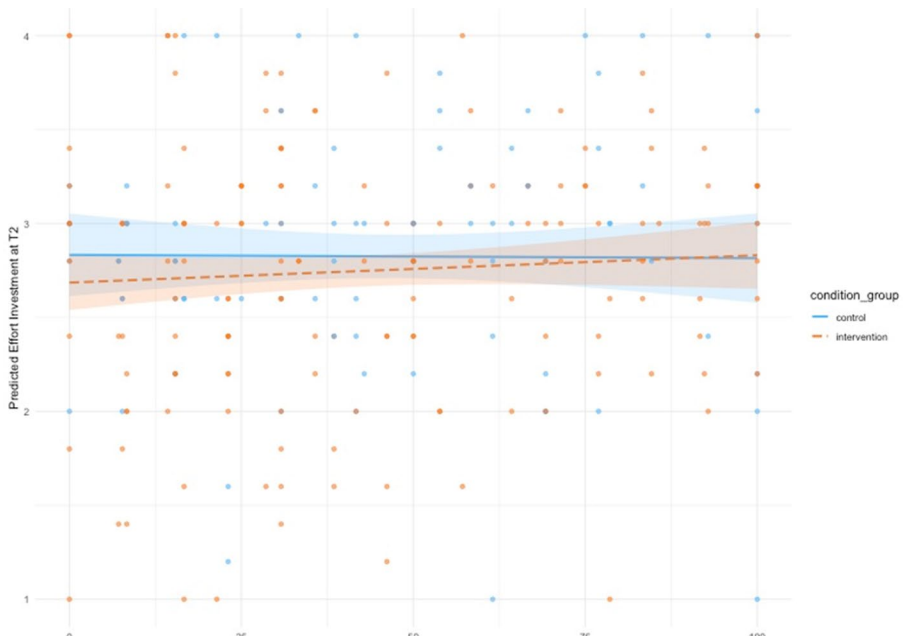


Fig. 18 Interaction of condition and meaningful interactions with the GPTs regarding effort investment. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

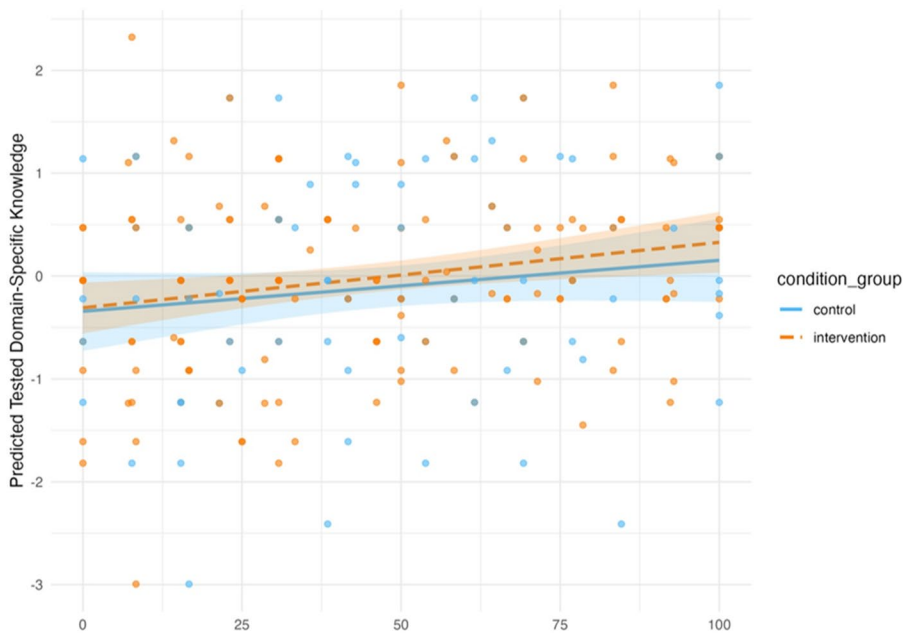


Fig. 19 Interaction of condition and meaningful interactions with the GPTs regarding domain-specific knowledge. *Note.* uv = utility value condition, cl = cognitive learning condition, co = control condition

Acknowledgements This research was supported by the Postdoctoral Academy of Education Sciences and Psychology of the Hector Research Institute of Education Sciences and Psychology, Tübingen, funded by the Baden-Württemberg Ministry of Science, Research, and the Arts and by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder. Lisa Bardach is supported by a Jacobs Foundation Research Fellowship, a Fellowship from the Elite Program for post docs by the Baden-Württemberg Foundation, and funding from research No P-EDU-23-28, which is co-funded by the European Union (the project “Breakthrough in Educational Research” No 10-044-P-0001) under the April 1, 2025 Agreement with the Research Council of Lithuania (RCOL, LMTLT) and the April 14, 2025 Joint Activity Agreement with Kaunas University of Technology.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability This study was preregistered at the Open Science Framework (OSF): <https://doi.org/10.17605/OSF.IO/VBSQ4>. The study’s data and Supplementary Material can be accessed via our OSF project page <https://doi.org/10.17605/OSF.IO/Y6KFS>. We have no known conflicts of interest to disclose. The authors are responsible for the content of this publication.

Declarations

Ethical approval An ethics committee approved the study and the data collection. It confirmed that the procedures were aligned with ethical research standards with human subjects (date of approval: February 20, 2025, file number: AZ.: A2.5.4–405_hb). In addition, the study was approved by the Ministry of Education, Youth, and Sports of the State of Baden-Württemberg. To ensure the absence of spelling errors and to obtain suggestions for more suitable phrasing, we used Grammarly (v1.114.5.0) and ChatGPT-5.2 (v1.2025.343) for text editing. ChatGPT was also used to support parts of the coding process.

Conflict of interest We have no known conflicts of interest to disclose. The authors are responsible for the content of this publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73(3), 277–320. <https://doi.org/10.3102/00346543073003277>
- Angrist, J. D., & Pischke, J. S. (2015). *Mastering 'metrics. The path from cause to effect*. Princeton University Press.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. <https://doi.org/10.1080/01621459.1996.10476902>
- Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist*, 50(1), 84–94. <https://doi.org/10.1080/00461520.2015.1004069>

- Azevedo, R., & Aleven, V. (Eds.). (2013). *International handbook of metacognition and learning technologies* (Vol. 28). Springer New York. <https://doi.org/10.1007/978-1-4419-5546-3>
- Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., Cloude, E., Dever, D., Wiedbusch, M., Wort, F., & Cerezo, R. (2022). Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Frontiers in Psychology*, 13, Article 813632. <https://doi.org/10.3389/fpsyg.2022.813632>
- Bardach, L., Oczlon, S., Pietschnig, J., & Lüftenegger, M. (2020). Has achievement goal theory been right? A meta-analysis of the relation between goal structures and personal achievement goals. *Journal of Educational Psychology*, 112(6), 1197–1220. <https://doi.org/10.1037/edu0000419>
- Bardach, L., Kalinowski, R., & Bailey, D. H. (2025). Differentiation in cognitive abilities beyond g: The emergence of domain-specific variance in childhood. *Psychological Science*, 36(3), 168–183. <https://doi.org/10.1177/09567976251321382>
- Bastian, A., Buchholtz, N., & Kaiser, G. (2025). *Using AI chatbots to facilitate mathematics preservice teachers' noticing skills*. <https://doi.org/10.13140/RG.2.2.1986.24004>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bernacki, M. L. (2023). Development, sustainment, and scaling of self-regulated learning analytics: Prediction modeling and digital student success initiatives in university contexts. In D. Glick, J. Bergin, & C. Chang (Eds.), *Advances in Educational Technologies and Instructional Design* (pp. 255–281). IGI Global. <https://doi.org/10.4018/978-1-6684-6500-4.ch012>
- Bernacki, M. L., Aguilar, Anita, C., & Byrnes, J. P. (2011). Self-regulated learning and technology-enhanced learning environments: An opportunity-propensity analysis. In G. Dettori & D. Persico (Eds.), *Fostering Self-Regulated Learning through ICT* (pp. 1–26). IGI Global Scientific Publishing. <https://doi.org/10.4018/978-1-61692-901-5>
- Boekaerts, M. (1996). Self-regulated learning at the junction of cognition and motivation. *European Psychologist*, 1(2), 100–112. <https://doi.org/10.1027/1016-9040.1.2.100>
- Boekaerts, M. (2007). Self-regulation and effort investment. In W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology* (pp. 345–377). John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470147658.chpsy0409>
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology*, 54(2), 199–231. <https://doi.org/10.1111/j.1464-0597.2005.00205.x>
- Breitwieser, J., Neubauer, A. B., Schmiedek, F., & Brod, G. (2022). Self-regulation prompts promote the achievement of learning goals – But only briefly: Uncovering hidden dynamics in the effects of a psychological intervention. *Learning and Instruction*, 80, Article 101560. <https://doi.org/10.1016/j.learninstruc.2021.101560>
- Brod, G. (2021). Generative learning: Which strategies for what age? *Educational Psychology Review*, 33(4), 1295–1318. <https://doi.org/10.1007/s10648-020-09571-9>
- Brod, G. (2024). There are multiple paths to personalized education, and they should be combined. *Current Directions in Psychological Science*, 33(3), 153–158. <https://doi.org/10.1177/09637214241242459>
- Bühler, B., Hou, R., Bozkir, E., Goldberg, P., Gerjets, P., Trautwein, U., & Kasneci, E. (2023). Automated hand-raising detection in classroom videos: A view-invariant and occlusion-robust machine learning approach. In N. Wang, G. Rebollo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (Vol. 13916, pp. 102–113). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_9
- Bühler, B., Bozkir, E., Goldberg, P., Sümer, Ö., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2024). From the lab to the wild: Examining generalizability of video-based mind wandering detection. *International Journal of Artificial Intelligence in Education*, 35(2), 823–857. <https://doi.org/10.1007/s40593-024-00412-2>
- Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C., & Wadman, K. (2025). *How people use ChatGPT*. OpenAI. <https://openai.com/research/how-people-use-chatgpt>
- Chen, J. (2022). The effectiveness of self-regulated learning (SRL) interventions on L2 learning achievement, strategy employment and self-efficacy: A meta-analytic study. *Frontiers in Psychology*, 13, Article 1021101. <https://doi.org/10.3389/fpsyg.2022.1021101>

- Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, pp. 161–238). Lawrence Erlbaum Associates.
- Cloude, E. B., Azevedo, R., Winne, P. H., Biswas, G., & Jang, E. E. (2022). System design for using multimodal trace data in modeling self-regulated learning. *Frontiers in Education*, 7, Article 928632. <https://doi.org/10.3389/educ.2022.928632>
- Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M., & Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. *British Journal of Educational Technology*, 56(5), 1686–1709. <https://doi.org/10.1111/bjet.13587>
- Cuschieri, S. (2019). The CONSORT statement. *Saudi Journal of Anaesthesia*, 13(5), 27. https://doi.org/10.4103/sja.SJA_559_18
- Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. (2016). Die deutsche Version des Big Five Inventory 2 (BFI-2). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. <https://doi.org/10.6102/ZIS247>
- De Mooij, S., Lämsä, J., Lim, L., Aksela, O., Athavale, S., Bistolfi, I., Jin, F., Li, T., Azevedo, R., Bannert, M., Gašević, 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(2), Article 54. <https://doi.org/10.1007/s10648-025-10028-0>
- Dent, A. L., & Koelka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(3), 425–474. <https://doi.org/10.1007/s10648-015-9320-8>
- Deutsche Akademie Der Naturforscher Leopoldina (with MyCoRe Community) (2024). Förderung der Selbstregulationskompetenzen von Kindern und Jugendlichen in Kindertageseinrichtungen und Schulen. In *Schriftenreihe zur wissenschaftsbasierten Politikberatung: Stellungnahme*. MyCoRe Community. https://doi.org/10.26164/LEOPOLDINA_03_01157
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(3), 231–264. <https://doi.org/10.1007/s11409-008-9029-x>
- Dignath, C., & Büttner, G. (2018). Teachers' direct and indirect promotion of self-regulated learning in primary and secondary school mathematics classes – insights from video-based classroom observations and teacher interviews. *Metacognition and Learning*, 13(2), 127–157. <https://doi.org/10.1007/s11409-018-9181-x>
- Dignath, C., & Veenman, M. V. J. (2021). The role of direct strategy instruction and indirect activation of self-regulated learning—Evidence from classroom observation studies. *Educational Psychology Review*, 33(2), 489–533. <https://doi.org/10.1007/s10648-020-09534-0>
- Dignath, C., Buettner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? *Educational Research Review*, 3, 101–129. <https://doi.org/10.1016/j.edurev.2008.02.003>
- Donker, A. S., De Boer, H., Kostons, D., Dignath Van Ewijk, C. C., & Van Der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, 11, 1–26. <https://doi.org/10.1016/j.edurev.2013.11.002>
- Du, J., Hew, K. F., & Liu, L. (2023). What can online traces tell us about students' self-regulated learning? A systematic review of online trace data analysis. *Computers & Education*, 201, Article 104828. <https://doi.org/10.1016/j.compedu.2023.104828>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, Article 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Eccles, J. S., & Wigfield, A. (2023). Expectancy-value theory to situated expectancy-value theory: Reflections on the legacy of 40+ years of working together. *Motivation Science*, 9(1), 1–121. <https://doi.org/10.1037/mot0000275>
- Efkliides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, 46(1), 6–25. <https://doi.org/10.1080/00461520.2011.538645>

- Elliot, A. J., & McGregor, H. A. (2001). A 2×2 achievement goal framework. *Journal of Personality and Social Psychology*, 80(3), 501–519. <https://doi.org/10.1037/0022-3514.80.3.501>
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjjet.13544>
- Fiorella, L. (2022). Learning by explaining after pauses in video lectures: Are provided visuals a scaffold or a crutch? *Applied Cognitive Psychology*, 36(5), 1142–1149. <https://doi.org/10.1002/acp.3994>
- Fiorella, L. (2023). Making sense of generative learning. *Educational Psychology Review*, 35(2), Article 50. <https://doi.org/10.1007/s10648-023-09769-7>
- Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity: Eight learning strategies that promote understanding* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781107707085>
- Foster, N., & Piacentini, M. (2023). *Innovating assessments to measure and support complex skills* (1st ed.). Organization for Economic Cooperation & Development (OECD).
- Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, 86, 186–204. <https://doi.org/10.1016/j.cortex.2016.04.023>
- Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Scientific Reports*, 13(1), Article 15310. <https://doi.org/10.1038/s41598-023-42227-6>
- Fütterer, T., Hoch, E., & Dumont, H. (2024). Uncovering the relationship between technology-enhanced, adaptive teaching and situational interest in mathematics in a randomized controlled trial. <https://doi.org/10.21203/rs.3.rs-5238796/v1>
- Fütterer, T., Pardi, G., Ziegs, T., Flegr, S., Ring, M., Fabian, A., Pawlak, F., Steinmetz, T., Becker, S., Bertram, C., Schreiber, W., Meurers, D., Brahm, T., Führer, C., Kuhn, J., Schwarzer, S., Paravicini, W., Burde, J. P., Nuxoll, F., & Gerjets, P. (2026). Zwischen Praxisrelevanz und technologischer Dynamik: Forschungsmethodische Perspektiven auf genAI im Unterricht am Beispiel niveaudifferenzierter Lernaufgaben. *Zeitschrift für Erziehungswissenschaft*.
- Gaspard, H., Dicke, A.-L., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, 107(3), 663–677. <https://doi.org/10.1037/edu0000003>
- Gaspard, H., Dicke, A.-L., Flunger, B., Häfner, I., Brisson, B. M., Trautwein, U., & Nagengast, B. (2016). Side effects of motivational interventions? Effects of an intervention in math classrooms on motivation in verbal domains. *AERA Open*, 2(2). <https://doi.org/10.1177/2332858416649168>
- Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, 48, 67–84. <https://doi.org/10.1016/j.cedpsych.2016.09.003>
- Geng, Z., Zeng, B., & Guo, L. (2024). Associations between behavioral, cognitive, and emotional self-regulation and academic and social outcomes among Chinese children: A meta-analysis. *Educational Psychology Review*, 36(1), Article 4. <https://doi.org/10.1007/s10648-023-09840-3>
- Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and meta-cognitive processes while using computer-based learning environments. *Educational Psychologist*, 45(4), 203–209. <https://doi.org/10.1080/00461520.2010.515935>
- Greene, J. A., Costa, L.-J., Robertson, J., Pan, Y., & Deekens, V. M. (2010). Exploring relations among college students' prior knowledge, implicit theories of intelligence, and self-regulated learning in a hypermedia environment. *Computers & Education*, 55(3), 1027–1043. <https://doi.org/10.1016/j.compedu.2010.04.013>
- Han, I., Ji, H., Jin, S., & Choi, K. (2025a). Mobile-based artificial intelligence chatbot for self-regulated learning in a hybrid flipped classroom. *Journal of Computing in Higher Education*, 38(1), 285–309. <https://doi.org/10.1007/s12528-025-09434-8>
- Han, X., Peng, H., & Liu, M. (2025b). The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies. *Educational Research Review*, 48, 100714. <https://doi.org/10.1016/j.edurev.2025.100714>
- Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights From the Behavioral and Brain Sciences*, 3(2), 220–227. <https://doi.org/10.1177/2372732216655542>

- Hausmann, R. G. M., & Vanlehn, K. (2007). Explaining self-explaining: A contrast between content and generation. *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 417–424.
- Hecht, C. A., Grande, M. R., & Harackiewicz, J. M. (2021). The role of utility value in promoting interest development. *Motivation Science*, 7(1), 1–20. <https://doi.org/10.1037/mot0000182>
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. Bis 12. Klassen, Revision (KFT 4–12 + R)*. Beltz Test.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4
- Hilpert, J. C., Greene, J. A., & Bernacki, M. (2023). Leveraging complexity frameworks to refine theories of engagement: Advancing self-regulated learning in the age of artificial intelligence. *British Journal of Educational Technology*, 54(5), 1204–1221. <https://doi.org/10.1111/bjet.13340>
- Hoch, E., Scheiter, K., & Stalbovs, K. (2023). How to support learning with multimedia instruction: Implementation intentions help even when load is high. *British Journal of Psychology*, 114(2), 315–334. <https://doi.org/10.1111/bjop.12620>
- Holmes, J. (2007). Designing agents to support learning by explaining. *Computers & Education*, 48(4), 523–547. <https://doi.org/10.1016/j.compedu.2005.02.007>
- Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence*, 5, Article 100165. <https://doi.org/10.1016/j.caeai.2023.100165>
- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, 326(5958), 1410–1412. <https://doi.org/10.1126/science.1177067>
- Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010a). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, 102(4), 880–895. <https://doi.org/10.1037/a0019506>
- Hulleman, C. S., Schragar, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010b). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, 136(3), 422–449. <https://doi.org/10.1037/a0018947>
- Jacob, L., Lachner, A., & Scheiter, K. (2020). Learning by explaining orally or in written form? Text complexity matters. *Learning and Instruction*, 68, Article 101344. <https://doi.org/10.1016/j.learninstruc.2020.101344>
- Jaekel, A. -K., Göllner, R., & Trautwein, U. (2020). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject: Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology*, 113(4), 770–783. <https://doi.org/10.1037/edu0000488>
- Jansen, R. S., Van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review*, 28, Article 100292. <https://doi.org/10.1016/j.edurev.2019.100292>
- Järvelä, S., Zhao, G., Nguyen, A., & Chen, H. (2025). Hybrid intelligence: Human–<Emphasis Type="SmallCaps">AI</Emphasis> coevolution and learning. *British Journal of Educational Technology*, 56(2), 455–468. <https://doi.org/10.1111/bjet.13560>
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kistner, S., Rakoczy, K., Otto, B., Dignath-van Ewijk, C., Büttner, G., & Klieme, E. (2010). Promotion of self-regulated learning in classrooms: Investigating frequency, quality, and consequences for student performance. *Metacognition and Learning*, 5(2), 157–171. <https://doi.org/10.1007/s11409-010-9055-3>
- Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. Springer. <https://doi.org/10.1007/978-0-387-77318-6>

- Klingsieck, K. B. (2018). Kurz und knapp – die Kurzskaala des Fragebogens „Lernstrategien im Studium“ (LIST). *Zeitschrift für Pädagogische Psychologie*, 32(4), 249–259. <https://doi.org/10.1024/1010-0652/a000230>
- Knogler, M., Harackiewicz, J. M., Gegenfurtner, A., & Lewalter, D. (2015). How situational is situational interest? Investigating the longitudinal structure of situational interest. *Contemporary Educational Psychology*, 43, 39–50. <https://doi.org/10.1016/j.cedpsych.2015.08.004>
- Kong, S.-C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K–12 settings. *IEEE Transactions on Learning Technologies*, 17, 1562–1573. <https://doi.org/10.1109/9/TLT.2024.3392830>
- Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., Kasneci, E., Kasneci, G., Kuhr, T., Kutyniok, G., Malone, S., Sailer, M., Schmidt, A., Stadler, M., Weller, J., & Kuhn, J. (2025). On opportunities and challenges of large multimodal foundation models in education. *Npj Science of Learning*, 10(1), Article 11. <https://doi.org/10.1038/s41539-025-00301-w>
- Kühl, T., Scheiter, K., Gerjets, P., & Gemballa, S. (2011). Can differences in learning strategies explain the benefits of learning from static and dynamic visualizations? *Computers & Education*, 56(1), 176–187. <https://doi.org/10.1016/j.compedu.2010.08.008>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. <https://doi.org/10.3102/0034654315581420>
- Lachner, A., Jacob, L., & Hoogerheide, V. (2021). Learning by writing explanations: Is explaining to a fictitious student more effective than self-explaining? *Learning and Instruction*, 74, Article 101438. <https://doi.org/10.1016/j.learninstruc.2020.101438>
- Lachner, A., Hoogerheide, V., van Gog, T., & Renkl, A. (2022). Learning-by-teaching without audience presence or interaction: When and why does it work? *Educational Psychology Review*, 34(2), Article 2. <https://doi.org/10.1007/s10648-021-09643-4>
- Lan, M., & Zhou, X. (2025). A qualitative systematic review on AI empowered self-regulated learning in higher education. *NPJ Science of Learning*, 10(1), Article 21. <https://doi.org/10.1038/s41539-025-00319-0>
- Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, 86(2), 602–640. <https://doi.org/10.3102/0034654315617832>
- Lenth, R. V. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version R package version 1.10.0) [Computer software]. <https://CRAN.R-project.org/package=emmeans>
- Lightbown, P., & Spada, N. M. (2013). *How languages are learned* (4th ed.). Oxford University Press.
- Lim, L., Bannert, M., Van Der Graaf, J., Singh, S., Fan, Y., Surendranair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior*, 139, Article 107547. <https://doi.org/10.1016/j.chb.2022.107547>
- Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, 70(4), 647–671. <https://doi.org/10.1177/0013164409355699>
- Makransky, G., Shiwalia, B. M., Herlau, T., & Blurton, S. (2025). Beyond the “Wow” factor: Using generative AI for increasing generative sense-making. *Educational Psychology Review*, 37(3), Article 60. <https://doi.org/10.1007/s10648-025-10039-x>
- Marsh, H. W. (1990a). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review*, 2(2), 77–172. <https://doi.org/10.1007/BF01322177>
- Marsh, H. W. (1990b). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, 82(4), 623–636. <https://doi.org/10.1037/0022-0663.82.4.623>
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6, Article 100199. <https://doi.org/10.1016/j.caeai.2023.100199>
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100. <https://doi.org/10.1006/cogp.1999.0734>
- Molenaar, I. (2022a). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. *Computers and Education: Artificial Intelligence*, 3, Article 100070. <https://doi.org/10.1016/j.caeai.2022a.100070>

- Molenaar, I. (2022b). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
- Molenaar, I., Mooij, S. D., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139, Article 107540. <https://doi.org/10.1016/j.chb.2022.107540>
- Möller, J. (2024). Ten years of dimensional comparison theory: On the development of a theory from educational psychology. *Educational Psychology Review*, 36(3), Article 82. <https://doi.org/10.1007/s10648-024-09918-6>
- Morehead, K., Dunlosky, J., Rawson, K. A., Bishop, M., & Pyc, M. A. (2018). Does mediator use contribute to the spacing effect for cued recall? Critical tests of the mediator hypothesis. *Memory*, 26(4), 535–546. <https://doi.org/10.1080/09658211.2017.1381266>
- Mousavinasab, E., Zarifsanaiy, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1), 142–163. <https://doi.org/10.1080/10494820.2018.1558257>
- Nagengast, B., Brisson, B. M., Hulleman, C. S., Gaspard, H., Häfner, I., & Trautwein, U. (2018). Learning more from educational intervention studies: Estimating complier average causal effects in a relevance intervention. *The Journal of Experimental Education*, 86(1), 105–123. <https://doi.org/10.1080/00220973.2017.1289359>
- Nguyen, H., Nguyen, V., Ludovise, S., & Santagata, R. (2025). Value-sensitive design of chatbots in environmental education: Supporting identity, connectedness, well-being and sustainability. *British Journal of Educational Technology*, 56(4), 1370–1390. <https://doi.org/10.1111/bjet.13568>
- OECD (2023). *Teaching for the future: Global engagement, sustainability and digital skills*. OECD. <https://doi.org/10.1787/d6b3d234-en>
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341. <https://doi.org/10.1007/s10648-006-9029-9>
- Pintrich, P. R. (2005). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Elsevier.
- Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation. In *Development of Achievement Motivation* (pp. 249–284). Elsevier. <https://doi.org/10.1016/B978-012750053-9/50012-7>
- Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, 52(3), 275–300. <https://doi.org/10.1080/15391523.2020.1719943>
- Posit team (2025). *RStudio: Integrated Development for R* (Version 2025.5.1.513) [Computer software]. Posit Software, PBC. <http://www.rstudio.com/>
- Praetorius, A.-K., Jentsch, A., Luoto, J., Keller, S. D., & Fauth, B. (2025). Context in research on teaching quality: Bringing a complex idea into the spotlight. *School Effectiveness and School Improvement*, 36(2), 263–296. <https://doi.org/10.1080/09243453.2025.2482577>
- R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <http://www.R-project.org>
- Rammstedt, B., Grüning, D. J., & Lechner, C. M. (2024). Measuring growth mindset: Validation of a three-item and a single-item scale in adolescents and adults. *European Journal of Psychological Assessment*, 40(1), 84–95. <https://doi.org/10.1027/1015-5759/a000735>
- Renninger, K. A., & Hidi, S. E. (2022). Interest: A unique affective and cognitive motivational variable that develops. In *Advances in Motivation Science* (Vol. 9, pp. 179–239). Elsevier. <https://doi.org/10.1016/bs.adms.2021.12.004>
- Renninger, K. A., & Su, S. (2012). Interest and its development. In R. M. Ryan (Ed.), *The Oxford Handbook of Human Motivation* (1st ed., pp. 167–188). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195399820.013.0011>
- Roll, I., Baker, R. S. J. D., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, 23(4), 537–560. <https://doi.org/10.1080/10508406.2014.883977>
- Rosenzweig, E. Q., Wigfield, A., & Eccles, J. S. (2022). Beyond utility value interventions: The why, when, and how for next steps in expectancy-value intervention research. *Educational Psychologist*, 57(1), 11–30. <https://doi.org/10.1080/00461520.2021.1984242>

- Sakaki, M., Murayama, K., Frenzel, A. C., Goetz, T., Marsh, H. W., Lichtenfeld, S., & Pekrun, R. (2024). Developmental trajectories of achievement emotions in mathematics during adolescence. *Child Development*, 95(1), 276–295. <https://doi.org/10.1111/cdev.13996>
- Scheiter, K., & Gerjets, P. (2002). The impact of problem order: Sequencing problems as a strategy for improving one's performance. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society* (1st ed., pp. 798–803). Routledge. <https://doi.org/10.4324/9781315782379>
- Scheiter, K., Schubert, C., Gerjets, P., & Stalbovs, K. (2015). Does a strategy training foster students' ability to learn from multimedia? *The Journal of Experimental Education*, 83(2), 266–289. <https://doi.org/10.1080/00220973.2013.876603>
- Schunk, D. H., & Greene, J. A. (Eds.). (2018). *Handbook of self-regulation of learning and performance* (2nd ed.). Routledge.
- Schunk, D. H., & Zimmerman, B. J. (Eds.). (1998). *Self-regulated learning: From teaching to self-reflective practice*. Guilford Press.
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46(3), 407–441. <https://doi.org/10.3102/00346543046003407>
- Stalbovs, K., Scheiter, K., & Gerjets, P. (2015). Implementation intentions during multimedia learning: Using if-then plans to facilitate cognitive processing. *Learning and Instruction*, 35, 1–15. <https://doi.org/10.1016/j.learninstruc.2014.09.002>
- Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, 11(1). <https://doi.org/10.1186/s40561-024-00354-1>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tinajero, C., Mayo, M. E., Villar, E., & Martínez-López, Z. (2024). Classic and modern models of self-regulated learning: Integrative and componential analysis. *Frontiers in Psychology*, 15, Article 1307574. <https://doi.org/10.3389/fpsyg.2024.1307574>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Vodafone Stiftung Deutschland gGmbH (2024). *Pioniere des Wandels. Wie Schüler:innen KI im Unterricht nutzen möchten*. <https://www.vodafone-stiftung.de/wp-content/uploads/2024/03/Pioniere-des-Wandels-wie-Schueler-innen-KI-im-Unterricht-nutzen-wollen-Jugendstudie-der-VS-2024.pdf>
- Wang, Y., & Sperling, R. A. (2020). Characteristics of effective self-regulated learning interventions in mathematics classrooms: A systematic review. *Frontiers in Education*, 5, Article 58. <https://doi.org/10.3389/educ.2020.00058>
- What Works Clearinghouse (2022). *Procedures and standards handbook, version 5.0* (Institute of Education Sciences, Ed.). U.S. Department of Education. https://ies.ed.gov/ncee/WWC/Docs/referencerources/Final_WWC-HandbookVer5_0-0-508.pdf
- Winne, P. H. (1995). Self-regulation is ubiquitous but its forms vary with knowledge. *Educational Psychologist*, 30(4), 223–228. https://doi.org/10.1207/s15326985ep3004_9
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Lawrence Erlbaum Associates.
- Wortha, F. D. (2021). *Beyond self-regulated learning: Integrating approaches to self-regulation in education*. Universität Tübingen.
- Wortha, F., Gerjets, P., Azevedo, R., Bernacki, M., Brucker, B., Bühler, B., Ehls, A.-C., Kasneci, E., Miyake, A., Murayama, K., Nagengast, B., Roberts, B., Tibus, M. & Trautwein, U. (2026). *A conceptual framework to integrate research programs on self-regulation in education*. Educational Psychology Review.
- 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. <https://doi.org/10.1111/bjet.13334>
- Xu, Z., Zhao, Y., Zhang, B., Liew, J., & Kogut, A. (2023). A meta-analysis of the efficacy of self-regulated learning interventions on academic achievement in online and blended environments in K-12 and higher education. *Behaviour & Information Technology*, 42(16), 2911–2931. <https://doi.org/10.1080/0144929X.2022.2151935>

- Xu, X., Qiao, L., Cheng, N., Liu, H., & Zhao, W. (2025). Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *British Journal of Educational Technology*(5). <https://doi.org/10.1111/bjet.13599>
- Yan, L., Greiff, S., Lodge, J. M., & Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology*, 4(7), 435–436. <https://doi.org/10.1038/s44159-025-00467-5>
- Yang, J. W., Choe, K., Lim, J., & Park, J. (2025). Asking questions matters: Comparing the effect of learning through dialogue with peers, ChatGPT, and intelligent tutoring system. *Interactive Learning Environments*, 33(10), 5870–5886. <https://doi.org/10.1080/10494820.2025.2488984>
- Zhu, W., Wang, F., Mayer, R. E., & Liu, T. (2024). Effects of explaining a science lesson to others or to oneself: A cognitive neuroscience approach. *Learning and Instruction*, 91, Article 101897. <https://doi.org/10.1016/j.learninstruc.2024.101897>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Tim Fütterer¹  · Lisa Bardach^{2,3}  · Jochen Kuhn⁴  · Stefan Daniel Keller⁵  · Peter Gerjets⁶ 

✉ Tim Fütterer
tim.fuetterer@uni-tuebingen.de

¹ University of Tübingen, Tübingen, Germany

² University of Giessen, Giessen, Germany

³ Kaunas University of Technology, Kaunas, Lithuania

⁴ Ludwig-Maximilians-Universität München, Munich, Germany

⁵ Zurich University of Teacher Education, Zurich, Switzerland

⁶ Leibniz-Institut für Wissensmedien, Tübingen, Germany