Computers in Human Behavior: Artificial Humans 9 (2026) 100378

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

How does collaborative prompting behaviour develop and relate to effectiveness in problem-solving with Generative Artificial Intelligence?

Athanasios Christopoulos [a,*], Stylianos Mystakidis [b], Isidoros Perikos [c,d], Aikaterini Bourazeri [e], Andria Michael [f,g], Mikko-Jussi Laakso [a]

[a] Turku Research Institute for Learning Analytics, Faculty of Science, University of Turku, Turku, FI-20014, Finland
[b] School of Natural Sciences, University of Patras, Rio, GR-26504, Greece
[c] Computer Engineering and Informatics Department, University of Patras, Rio, GR-26504, Greece
[d] Computer Technology Institute and Press "Diophantus", Patras, GR-26504, Greece
[e] School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom
[f] InSPIRE Research Centre, University of Central Lancashire Cyprus, 12–14 University Avenue, Pyla, 7080, Cyprus
[g] Research Centre for Migration, Diaspora, and Exile, University of Central Lancashire, Preston, PR1 2HE, United Kingdom

ARTICLE INFO

Keywords:
Generative Artificial Intelligence
Collaborative problem-solving
Prompting behaviour
Prompting self-efficacy

ABSTRACT

Research on Generative Artificial Intelligence (GenAI) in education has concentrated on user attitudes and adoption intentions but less is known about how groups actually interact with such tools while solving complex problems. The present study treats prompting as observable, time-resolved, behaviour and investigates how a group's prompting develops and relates to task effectiveness over a collaborative session. Secondary (N = 114) and Higher (N = 57) Education students, working in self-formed groups (N = 45), completed a competitive mystery-solving task with unrestricted access to GenAI tools of their choice. Logged group prompts were coded with a scheme adapted from Bloom's Digital Taxonomy that distinguishes lower-order (Exploration and Action) from higher-order (Creation and Metacognition) cognitive behaviours. Groups produced a modest majority of higher-order prompts and shifted progressively from exploratory toward metacognitive prompting as sessions advanced. A higher proportion of higher-order prompting was associated with greater effectiveness which held even when the total prompt volume was controlled. Prior experience, general attitudes toward AI, and self-reported verification behaviour showed no reliable association with effectiveness. Perceived usefulness and prompting self-efficacy co-varied with success but are interpreted as concurrent correlates. Foregrounding what groups did with the tool and how that behaviour developed over time offers a view on collaborative human-AI interaction that adoption-focused measures alone cannot provide.

1. Introduction

Generative Artificial Intelligence (GenAI) is increasingly used in educational and professional contexts with research pointing to gains in productivity, personalised learning, and complex problem-solving (Dwivedi et al., 2023; Kasneci et al., 2023). However, the ease of access to Large Language Models—such as ChatGPT, Gemini, Perplexity—has also raised concerns about how users interact with such systems and what factors shape the resulting cognitive outcomes (Eysenbach, 2023; Lee & Cha, 2025).

According to Knoth et al. (2024), the potential of GenAI to produce meaningful outcomes depends on users' ability to utilise the tool effectively. The authors also mention that many users—especially novices—struggle to formulate prompts that elicit accurate, relevant, and high-quality responses (Chiu, 2024a; Knoth et al., 2024) with recent reviews (Barrot, 2023; Shi & Aryadoust, 2024) confirming that the quality of Human-AI collaboration depends heavily on prompt formulation and call for systematic research on prompting behaviours.

Generating meaningful prompts differs fundamentally from traditional information retrieval where keyword searches suffice. Users who interact with GenAI tools must engage in communicative acts that span from simple factual queries to complex requests for synthesis, evaluation, and creative generation (Qian, 2025). In a sense, the cognitive demands of prompting can be paralleled to Bloom's higher-order

* Corresponding author.
E-mail addresses: atchri@utu.fi (A. Christopoulos), smyst@upatras.gr (S. Mystakidis), perikos@ceid.upatras.gr (I. Perikos), a.bourazeri@essex.ac.uk (A. Bourazeri), AMichael4@uclan.ac.uk (A. Michael), milaak@utu.fi (M.-J. Laakso).

https://doi.org/10.1016/j.chbah.2026.100378
Received 22 June 2026; Received in revised form 3 August 2026; Accepted 16 August 2026
Available online 22 August 2026

thinking skills wherein prompt quality may reflect the user's underlying cognitive engagement (Gonsalves, 2026; Ng et al., 2021). Under this notion, it can be argued that users who passively consume AI-generated content may fail to develop critical evaluation skills whereas, those who engage in iterative dialogue, may extract greater value from the interaction while developing their metacognitive skills.

Reviews of ChatGPT use in Higher Education indicate widespread student adoption for academic purposes alongside limited confidence in prompting skills (Ansari et al., 2024). Similarly, studies of actual prompting behaviour reveal wide variation in quality as many users default to simple factual queries instead of engaging in iterative, evaluative, exchanges that characterise expert use (Imran & Almusharraf, 2023; Mahapatra, 2024). Nevertheless, the factors that predict prompting effectiveness are still not well understood with reviews indicating that the cognitive dimensions of prompt construction have received far less attention than surface-level features—such as prompt length or specificity (Knoth et al., 2024; Qian, 2025). The observed discrepancy between access and effective use raises questions about the factors that determine whether Human-AI collaboration succeeds—especially in educational settings—where both the learning process and the task outcome matter.

Beyond the individual-level concerns, a further gap exists at the level of the group. Prompting has so far been studied mainly as the activity of individuals working with GenAI on their own (Knoth et al., 2024; Zamfirescu-Pereira et al., 2023) whereas, collaborative settings, in which a group jointly formulates, evaluates, and refines its queries, remain largely unexplored (Kim et al., 2024; Perifanou & Economides, 2025). The present study addresses this inadequacy of the literature by treating a group's prompting as observable behaviour that develops over the course of a task and by asking whether its cognitive quality relates to the effectiveness of the collaboration.

2. Background

2.1. Theoretical framework

Two complementary theoretical frameworks are often employed to explain the patterns regarding the differential use of AI: *1)* the Technology Acceptance Model (TAM) and *2)* the Automation Complacency Theory (ACT). With respect to the former, user acceptance of Information Technology is suggested to be determined primarily by two cognitive beliefs: *a)* Perceived Usefulness (PU), which is defined as the degree to which a person believes that using a particular system would enhance their performance, and *b)* Perceived Ease of Use (PEU), which refers to the degree to which a person believes that using a system would be free of effort. Subsequent studies show that PU correlates more strongly with usage than PEU and that PEU may act as a causal antecedent of PU instead of a parallel determinant (Kamarova et al., 2025; Şahin & Yıldız, 2024). The model has also been extended to include social influence and cognitive-instrumental processes thus suggesting that technology acceptance operates through both affective and rational pathways (Siu & White, 2025; Sun & Luo, 2024).

Evidence from GenAI contexts confirms the model's applicability for predicting usage behaviour and adoption intentions. For instance, Dahri et al. (2024) found that PU predicted ChatGPT adoption intentions among students engaged in metacognitive self-regulated learning with effect sizes comparable to those observed in studies of more traditional educational technologies. Likewise, the cross-cultural work by Budhathoki et al. (2024) demonstrated that the TAM constructs operate similarly across diverse cultural and educational contexts; a finding that points to generalisable mechanisms underlying GenAI acceptance. In the same vein, Barakat et al. (2025) reported that university educators' intentions to integrate ChatGPT into teaching were strongly predicted by PU with instructors who believed AI would enhance their pedagogical effectiveness showing greater adoption intentions. The aforementioned studies establish that acceptance beliefs matter for GenAI engagement but they address adoption intentions instead of performance outcomes.

Even though TAM explains adoption, it does not address what happens after users accept the technology. For the post-adoption stage of Human-AI interaction, ACT (Parasuraman & Manzey, 2010) provides a complementary lens. Complacency occurs when users defer to a technology they perceive as more proficient than themselves and, as such, fail to attend to the possibility that it may err. Automation bias represents a more extreme manifestation of the phenomenon wherein users actively prefer system signals over contradictory information from other sources, including their own senses. Parasuraman and Manzey (2010) showed that such phenomena arise from the interaction of personal, situational, and automation-related factors with attention allocation playing a central role.

The theory builds on a longer tradition of research on human interaction with automation. Parasuraman and Riley (1997) distinguished the use, misuse, and disuse of automated systems with misuse denoting over-reliance that produces monitoring failures and biased decision-making. Subsequent experimental works have sharpened the picture further. Participants supported by a highly, though not perfectly, reliable automated aid committed both omission errors—i.e., failing to notice events the aid did not flag—and commission errors—i.e., following its advice even when contradictory information was readily available (Skitka et al., 1999). Recent evidence indicates that similar dynamics extend to GenAI. Surveying knowledge workers, Lee et al. (2025) found that greater confidence in GenAI was associated with less critical thinking whereas greater confidence in one's own abilities was associated with more, thus suggesting that calibrating trust and reliance to the system's actual reliability remains a central challenge in conversational systems. As such, confidence in one's own ability to direct such systems—hereinto referred to as 'prompting self-efficacy'—may accordingly matter alongside confidence in the system itself.

In educational contexts, GenAI accuracy varies by domain and task type. Verification and critical evaluation of AI outputs may, therefore, distinguish effective from ineffective collaboration (Ivanov & Soliman, 2023; van Dis et al., 2023). However, the question of whether prior AI experience facilitates or inhibits verification behaviours remains empirically unresolved with some studies suggesting that experienced users develop appropriate skepticism whereas others indicate that familiarity breeds complacency.

The present study utilises both frameworks for their complementary functions. TAM motivates the acceptance-related measures and the expectation that beliefs about the tool's utility relate to task performance. ACT motivates the experience, trust, and verification measures, together with the corresponding expectations about over-reliance; jointly, the two frameworks also underpin the exploratory examination of how individual-level characteristics relate to group effectiveness. Neither framework, however, specifies what effective engagement with GenAI looks like at the level of observable behaviour. For this purpose, the study draws on Bloom-based descriptions of AI-supported learning, which supply the cognitive vocabulary for classifying prompts, and on the Human-AI collaboration literature, which links such behaviour to the effectiveness of the collaboration.

2.2. Prompting behaviours and cognitive engagement

Artificial Intelligence literacy denotes humans' socio-technical competencies regarding such tools; a construct comprising multiple proficiency dimensions that differ between non-expert and expert users and is assessed multidimensionally (i.e., not as a single skill) (Pinski & Benlian, 2024). Evidence from Human-Computer Interaction research points in the same direction. Non-expert users have been observed to design prompts opportunistically instead of systematically (Zamfirescu-Pereira et al., 2023) thus, indicating that prompting behaviour is shaped jointly by dispositions and by the task at hand. To this end, the study of Wyszynski et al. (2026) across five problem-solving scenarios found that trait satisficing reduced, whereas self-rated

prompting competence increased the pursuit of better outputs, with the balance shifting between task types.

Recent research has also examined prompting strategies as the key mechanism through which users shape AI collaboration outcomes (e.g., Chiu, 2024b; Knoth et al., 2024; Qian, 2025). Gonsalves (2026) demonstrated that prompting behaviours influence learning outcomes and further argues that, in order for AI-assisted learning to be effective, recursive engagement across cognitive domains is required. In response to this observation, the author proposed a revised Bloom's taxonomy to incorporate AI-specific competencies, arguing that the traditional hierarchical models are not capable of capturing the iterative, non-linear, nature of learning with GenAI. Similarly, Knoth et al. (2024) found that certain aspects of AI literacy were related to prompt-engineering skill, consistent with treating prompting behaviours as observable indicators of underlying cognitive engagement with AI systems. The findings align with the broader field of educational research which demonstrates that higher-order thinking manifests in how learners interact with instructional resources.

Bloom-based taxonomies have recently been adapted to describe how learners engage with AI thus, providing a baseline for the classification logic adopted here. Hmoud and Shaqour (2024) proposed a six-level model that maps AI-supported activities onto the revised taxonomy in which collecting and organising resources occupy the lowest level, interpreting information a middle level, and evaluating and creating the two highest. In the same vein, Hui (2025) applied the revised taxonomy to pupils' chatbot interactions and found that Evaluating and Creating dominated the higher-order processes students displayed with reliance on lower-level processes increasing as tasks grew more complex. Building on these models, the present study distinguishes six prompt-level behaviours—Searching, Reading, Interpreting, Evaluating, Organising, and Curating—and groups them into a lower-order stage, Exploration and Action (E&A), and a higher-order stage, Creation and Metacognition (C&M); together they constitute the Learning Behaviour with Emerging Technologies (LBET) framework employed throughout.

It should be noted that, the stage labels are our own synthesis: 'Creation' reflects the synthetic character of Organising—structuring distributed evidence into a coherent conclusion, which is what solving a problem requires—and 'Metacognition' reflects the evaluative and reflective oversight inherent in Evaluating and in Curating one's use of the tool, in line with the placement of these processes at the upper levels of the source models. Interpreting is placed in the lower-order stage because explaining and making sense of given information sits below evaluation and synthesis in those models (Hmoud & Shaqour, 2024).

2.3. Human-AI collaboration

Research on Human-AI collaboration has also highlighted the importance of cognitive engagement in determining collaboration effectiveness. Järvelä et al. (2025) conceptualised 'hybrid intelligence' as the productive interplay between human cognitive flexibility and AI computational capacity. Central to their framework is that successful collaboration depends on users actively directing AI toward goals neither human nor machine could reach alone. The experimental study of Wang et al. (2024) in Physics education reinforces the aforementioned claim with the authors concluding that, even though Human-AI collaboration improved problem-solving performance when compared with unaided performance, students often treated ChatGPT as an answer-generating tool instead of a genuine collaborative partner, thus missing opportunities for deeper learning through interactive exchange. Understanding variation in effectiveness, therefore, rests on the distinction between passive AI use (i.e., treating the system as an oracle) and active collaboration (i.e., iterative dialogue that builds toward solutions).

In the present study, effective collaboration is defined as the extent to which a group directs, evaluates, and integrates the AI's contributions in the service of a joint task outcome. It is assessed through two complementary indicators: *1)* the effectiveness score the group achieves and *2)* the cognitive quality of the prompts through which the group engages the tool. The two are conceptually distinct but closely linked. Active versus passive use describes an orientation toward the tool whereas prompt-level coding captures the observable trace that orientation leaves. In simple words, a group that merely consumes AI output has little occasion to produce evaluative, organisational, or synthetic prompts. The proportion of higher-order prompting therefore serves as a behavioural indicator of active collaboration in the sense of appropriate reliance as the discriminating use of AI advice (Schemmer et al., 2023, pp. 410–422).

Beyond collaboration, the groups also competed against one another, by design. Competition of this kind is an established motivational device; leaderboards function as implicit goal-setting mechanisms that can raise performance to levels comparable with difficult assigned goals (Landers et al., 2017). Meta-analytic evidence shows that gamification benefits cognitive, motivational, and behavioural learning outcomes with the combination of competition and collaboration being particularly effective for the latter (Sailer & Homner, 2020). Because the competitive element operated between groups (i.e. competitors) only, it was expected to energise, instead of disturbing, cooperation within them.

2.4. The present study

The few available studies of group-level prompting are descriptive accounts of how teams share prompting duties when working with a chatbot (Perifanou & Economides, 2025) or indicate that introducing one can redirect interaction away from peers and toward the tool itself (Kim et al., 2024). Across the identified research streams, experience, acceptance beliefs, and prompting behaviour have mostly been examined in isolation (Knoth et al., 2024; Qian, 2025). For instance, TAM research focuses on adoption intentions instead of actual performance outcomes whereas, prompting studies, often lack systematic frameworks to categorise prompt quality. At the same time, Human-AI collaboration research rarely connects interaction patterns to the individual difference variables that might predict them. Therefore, it remains unclear how technology acceptance beliefs, prior experience, and prompting behaviours combine to influence collaborative task effectiveness. Similarly, prior work has examined prompting mostly through static snapshots, such as single prompts, aggregate counts, or post-hoc self-report (Knoth et al., 2024; Qian, 2025) but how the cognitive complexity of a group's prompting shifts as one task unfolds has received far less attention.

The present study examines collaborative problem-solving with GenAI tools by groups of Secondary and Higher-Education students through a descriptive, group-level lens, by relating logged prompting behaviour to individual-level measures of experience, attitudes, and technology acceptance. The contribution is, accordingly, both empirical and methodological on three counts. First, the analysis uses logged group prompts produced in an ecologically valid task in which tool choice and prompting were unrestricted. Second, the study adapts Bloom's Digital Taxonomy (Hmoud & Shaqour, 2024; Hui, 2025) into a prompt-level coding scheme—the LBET framework—that locates each prompt on a continuum of cognitive complexity. Third, it charts the within-session development of prompting.

To structure the investigation the following Research Questions (RQs) were put forward. The first two concern prompting behaviour and were treated as confirmatory; the third is exploratory, examining how individual-level characteristics relate to a group-level outcome.

> **RQ1.** What kinds of prompts, along a continuum from information gathering to metacognition, do groups produce when solving a complex problem collaboratively with GenAI and how does the balance of lower- and higher-order prompting change over the course of a session?

RQ2. Is the proportion of higher-order prompting associated with group task effectiveness?

RQ3. How do prior GenAI experience, attitudes toward AI, technology acceptance beliefs, and verification behaviour relate to group task effectiveness?

To facilitate the examination of the proposed RQs, the following Research Hypotheses (*Hs*) were formed:

H1. Groups will demonstrate greater prevalence of C&M prompts compared to E&A prompts when solving complex collaborative tasks with GenAI.

Complex problem-solving tasks require synthesis, evaluation, and integration of information instead of retrieval (Gonsalves, 2026). Groups facing ill-structured problems must evaluate evidence quality, compare alternative explanations, and synthesise disparate information—activities that correspond to C&M instead of E&A prompts. Iterative refinement and critical evaluation of outputs recur in accounts of effective GenAI use (Knoth et al., 2024; Zamfirescu-Pereira et al., 2023) which further suggests that goal-directed collaboration will favour higher-order prompting.

H2. Higher proportions of C&M prompts will be associated with group task effectiveness after accounting for total prompt volume.

Groups that engage in more evaluative, organisational, and curatorial interactions with GenAI are expected to achieve better outcomes because C&M behaviours facilitate critical assessment of AI outputs, integration of information across multiple queries, and strategic application of AI-generated content to the task at hand (Järvelä et al., 2025; Wang et al., 2024).

H3. Prior GenAI experience will show a non-linear relationship with task effectiveness as a reflection of automation complacency.

We anticipate that moderate experience levels will be associated with optimal outcomes as, at low levels, insufficient prompting skills may limit effective use; at high levels, automation complacency may lead to over-reliance and reduced critical evaluation (Parasuraman & Manzey, 2010).

H4. PU will be associated with task effectiveness.

According to TAM, users who perceive greater usefulness engage more deeply and strategically with the technology. Recent extensions of TAM to GenAI contexts have confirmed that PU remains a significant predictor of both effective AI adoption and use (Budhathoki et al., 2024; Dahri et al., 2024).

H5. Verification behaviour—the tendency to check AI outputs—will be associated with task effectiveness.

Drawing on automation complacency research, which demonstrates that trust calibration requires active monitoring and cross-checking of automated outputs (Parasuraman & Manzey, 2010), we hypothesise that groups whose members verify the AI-generated information more frequently will achieve superior outcomes by avoiding errors stemming from uncritical acceptance of incorrect or incomplete AI-generated responses.

3. Materials and methods

3.1. Research design

A cross-sectional observational study was conducted, following the design typology of Creswell and Creswell (2018), to investigate how groups of students use GenAI for collaborative problem-solving during a

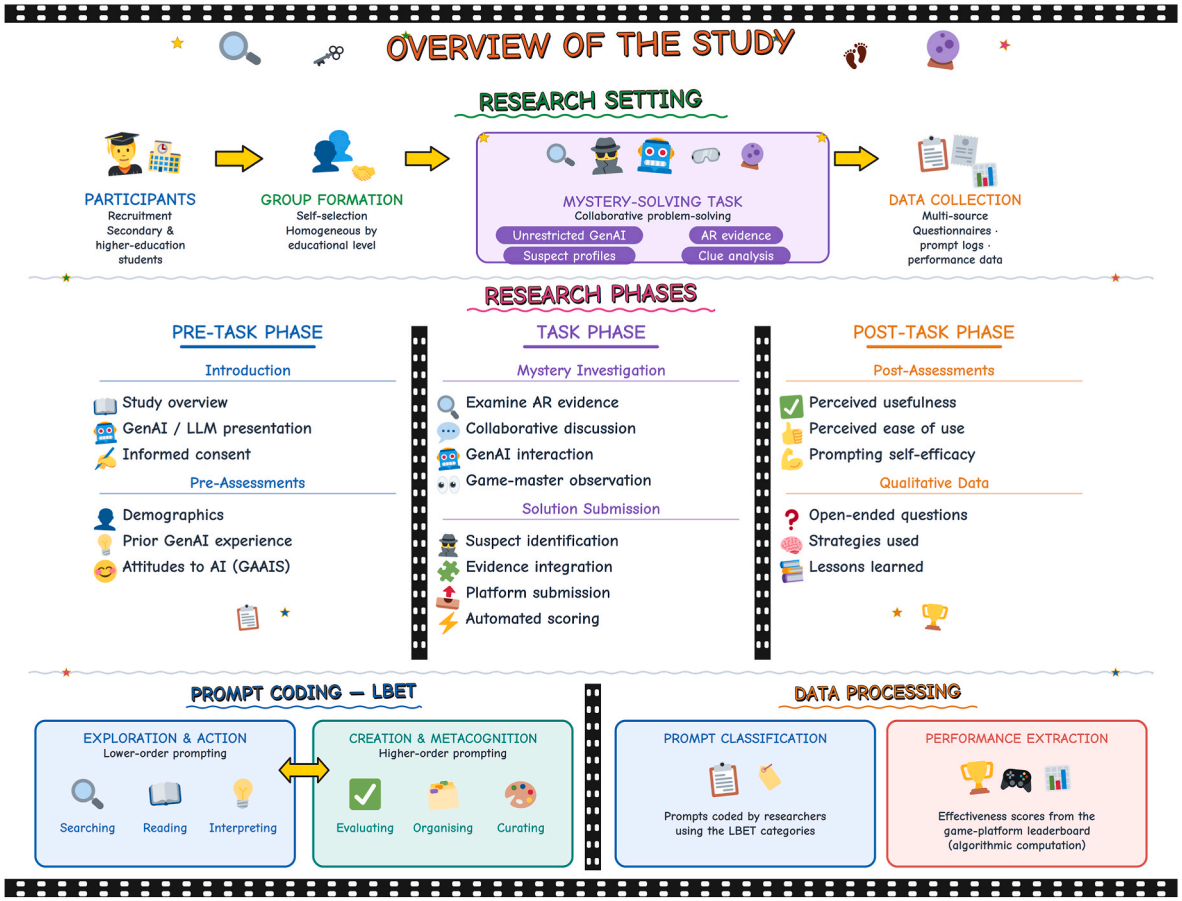

Fig. 1. Overview of the research design.

criminal-investigation mystery game in which the evidence was distributed through Augmented Reality (AR) markers in the physical environment (Fig. 1). The design and its measures are grounded in the two frameworks set out earlier—the TAM, which motivates the perceived-usefulness and ease-of-use measures, and the ACT, which motivates the prior-experience and verification measures. The study employed a single-session design, expected to last at most 90 min, with multiple sessions conducted to accommodate the full sample.

The observational design was used to document authentic group prompting as it unfolded instead of manipulating it. Prior experience and attitudes toward AI were collected before the task, along with two dispositional measures, the Cognitive Reflection Test and the Intolerance of Uncertainty Scale, which lie outside the present research questions and are not analysed here; prompting behaviour and effectiveness were recorded during and at the conclusion of the task; and technology acceptance (PU and PEU) together with Prompting Self-Efficacy (PSES) were collected immediately afterwards, since such judgements presuppose direct experience of the tool in the task at hand.

Each session began with a brief informational presentation about GenAI and Large Language Models, provided as an educational benefit to students. Participants then completed the pre-event psychometrics individually. Following pre-task measures, groups were formed and were introduced to the problem-solving task—a complex fictional mystery requiring resolution. An observer was assigned to each group in the role of Game Master (GM). Accordingly, groups worked collaboratively to analyse evidence, formulate prompts, evaluate AI-generated responses, and synthesise their conclusions. Collaboration occurred within each group whereas the competitive element operated between groups—a shared leaderboard ranked teams by their effectiveness scores as a means of providing task motivation without bearing on within-group cooperation. Upon completing the mystery, groups submitted their solution through the game platform which automatically calculated effectiveness scores. Participants then completed the post-event psychometrics individually. Prompt logs were copied over and coded at a later point using the proposed framework. Within these sessions, completion of the mystery task itself took between 15 and 61 min ($M = 32.38$, $SD = 12.19$). Participants received no monetary compensation for taking part though the top three performing teams received cinema tickets as a symbolic prize.

3.2. Participants

The sample comprised 171 participants organised into 45 groups. Participants were drawn from one Higher Education Institution and several private—supplementary education—centers that provide out-of-school academic support. The sample included 103 females (60.2%) and 68 males (39.8%) with ages ranging from 15 to 32 years ($M = 18.12$, $SD = 3.42$). Regarding the educational level, 114 participants (66.6%) were enrolled in secondary education, 49 (28.7%) were undergraduates, and 8 (4.7%) were postgraduate students. Academic backgrounds were diverse with secondary education students being divided across the Science and Technological Sciences track ($n = 41$), the Economics and Information Technology track ($n = 37$), the Humanities track ($n = 23$), and the Health and Life Sciences track (n=13). The university sample comprised students in computing and engineering fields.

Participants self-selected into groups of three to five members ($M = 3.80$) with group composition being homogeneous by educational level to control for potential confounds arising from disparate educational backgrounds within groups. Six groups combined undergraduate and postgraduate members. The distribution of group sizes was as follows: 11 groups of three members, 32 groups of four members, and 2 groups of five members. All participants provided informed consent and were free to withdraw at any point.

3.3. Materials and task

For the needs of the study, a collaborative digital mystery-solving game was developed; no participant had encountered the game before the sessions. The scenario presented a complex criminal investigation requiring analysis of forensic evidence, suspect testimonies, and timeline reconstruction. Groups investigated the death of a publishing executive found in his office building. The initial indications suggested suicide by poisoning but several anomalies pointed toward potential homicide. The case file included forensic reports, a detailed event chronology, and the profiles of five suspects, each with plausible motives and opportunities (Figs. 2 and 3).

To reach a conclusion, groups were required to examine 11 AR-enabled evidence pieces (Fig. 4) distributed across the physical environment, evaluate the suspect alibis, identify inconsistencies, and synthesise their findings to determine whether the death was suicide or murder and, if the latter, identify the perpetrator(s). The GenAI tools had no direct access to the case materials; groups conveyed evidence content into their prompts by typing or summarising it. Task difficulty was calibrated through pilot sessions with members of the research team.

The task was designed to require higher-order cognitive engagement in order to benefit from GenAI assistance. During the activity, participants were granted the freedom to use any GenAI tool of their choice. Restrictions were lifted to enhance ecological validity and capture authentic variation in tool choice and use. In practice the toolkit was largely homogeneous; every participant who reported prior GenAI experience (160 of the 171) named ChatGPT among the tools they used, Google Gemini was the second most common, and Perplexity was mentioned by a small minority, with no other tool reported at baseline.

During the sessions, the GenAI tool each team employed was recorded as a team-level variable: 18 teams worked with ChatGPT, 16 with Google Gemini, and 11 with Perplexity. All teams accessed their chosen tool through its web interface on their own devices. Following completion of the activity, the GMs recorded all the prompts submitted by the groups to the GenAI tools and rated their reasoning level. The complete environment is preserved as an interactive online demonstration and is available from the corresponding author upon request.

Mystery-solving effectiveness was computed algorithmically by the game platform and was based on four weighted components: solution accuracy, completion time, evidence correctly interpreted, and reasoning quality (Appendix A). The respective formula is presented below:

$$ME = 40S + 20(1 - T/90) + 20(E/11) + 20R$$

where S denotes the solution accuracy (correct suspect identification, 0–1), T denotes the completion time in minutes (capped at 90), E denotes the number of evidence pieces correctly interpreted (0–11), and R denotes the reasoning quality as rated by the GMs on a continuous scale (0–1).

An evidence piece was counted as correctly interpreted when the group's final solution drew from it the inference intended by the case design. The reasoning score was assigned by the GM through a structured observation rubric completed during the session; the rubric assessed systematic analysis of the evidence, cross-referencing of its individual pieces, and critical questioning of conclusions. The weighting reflects the design logic of the task: identifying the culprit is the criterial outcome and therefore carries the largest weight, while efficiency, evidentiary grounding, and reasoning quality contribute equally to the remainder. As any fixed weighting is to a degree conventional, the robustness of the focal association to alternative compositions is examined in the Results section. Effectiveness scores could range from 0 to 100; the observed values in the sample ranged from 55 to 96.

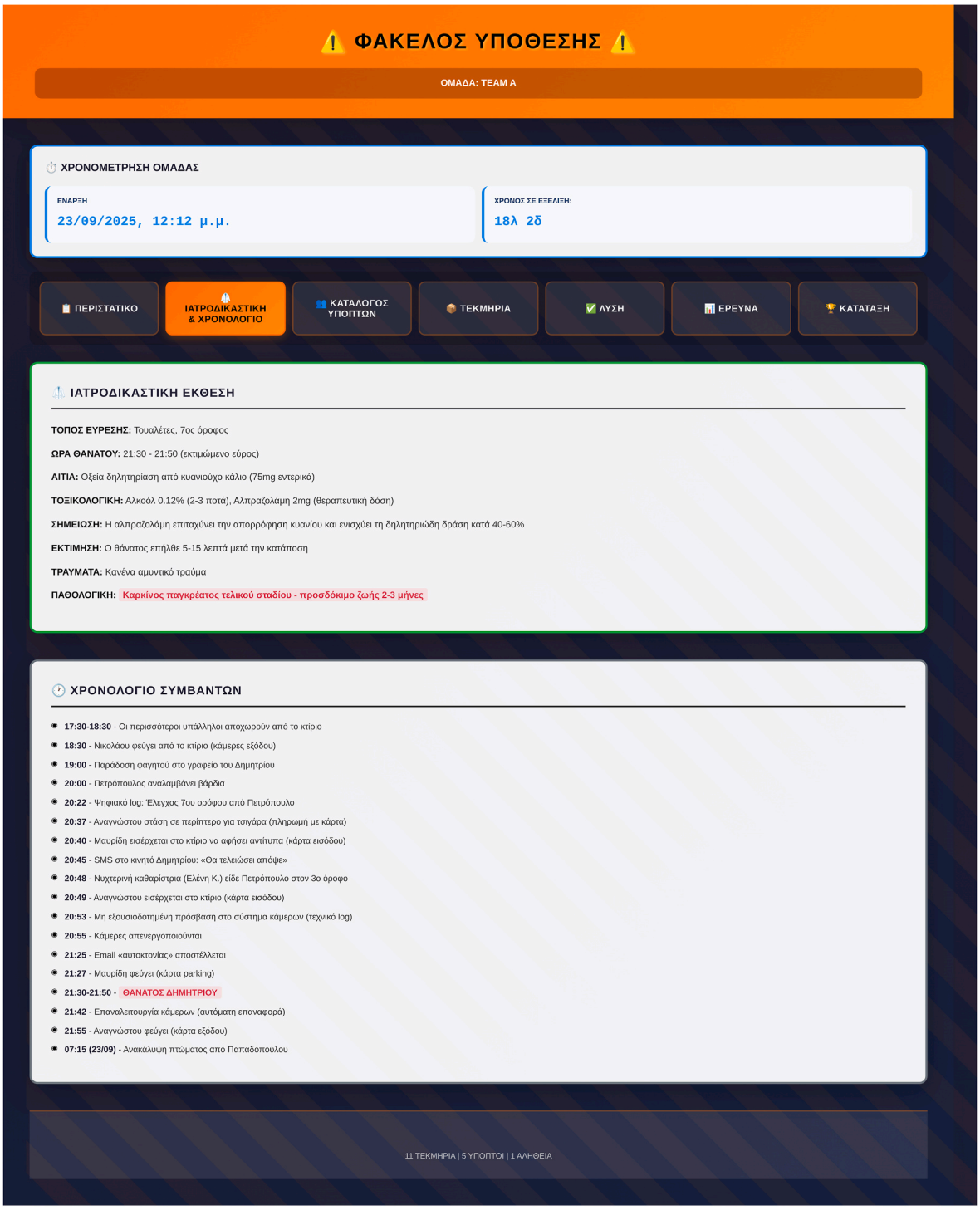

Fig. 2. Student-facing interface of the Mystery Crime Investigation environment (1/2).

3.4. Instruments

Participants first completed a demographics survey (Appendix B) which assessed gender, age, academic discipline, and educational level. A branched questionnaire then assessed GenAI experience (Appendix C) wherein *users* reported duration of use, frequency, tools employed, purposes, perceived capability at formulating prompts, trust in AI responses, and verification behaviours and *non-users* reported reasons for non-use, awareness of capabilities, concerns, and likelihood of future use. All participants rated the perceived importance of GenAI for their future careers and the adequacy of their education regarding AI use.

Attitudes toward AI were measured using the 8-item General Attitudes towards Artificial Intelligence Scale (GAAIS; Appendix D) (Schepman & Rodway, 2023) consisting of five positively worded items (e.g., "I am impressed by what Artificial Intelligence can do") and three negatively worded items (e.g., "I think Artificial Intelligence is dangerous"). Responses were recorded on a 5-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Internal consistency in the present sample was acceptable ($\alpha = .74$).

Technology acceptance was assessed with items adapted from the classic TAM scales (Davis, 1989), as applied to the GenAI context by Strzelecki (2024) (Appendix E). PU was measured with six items

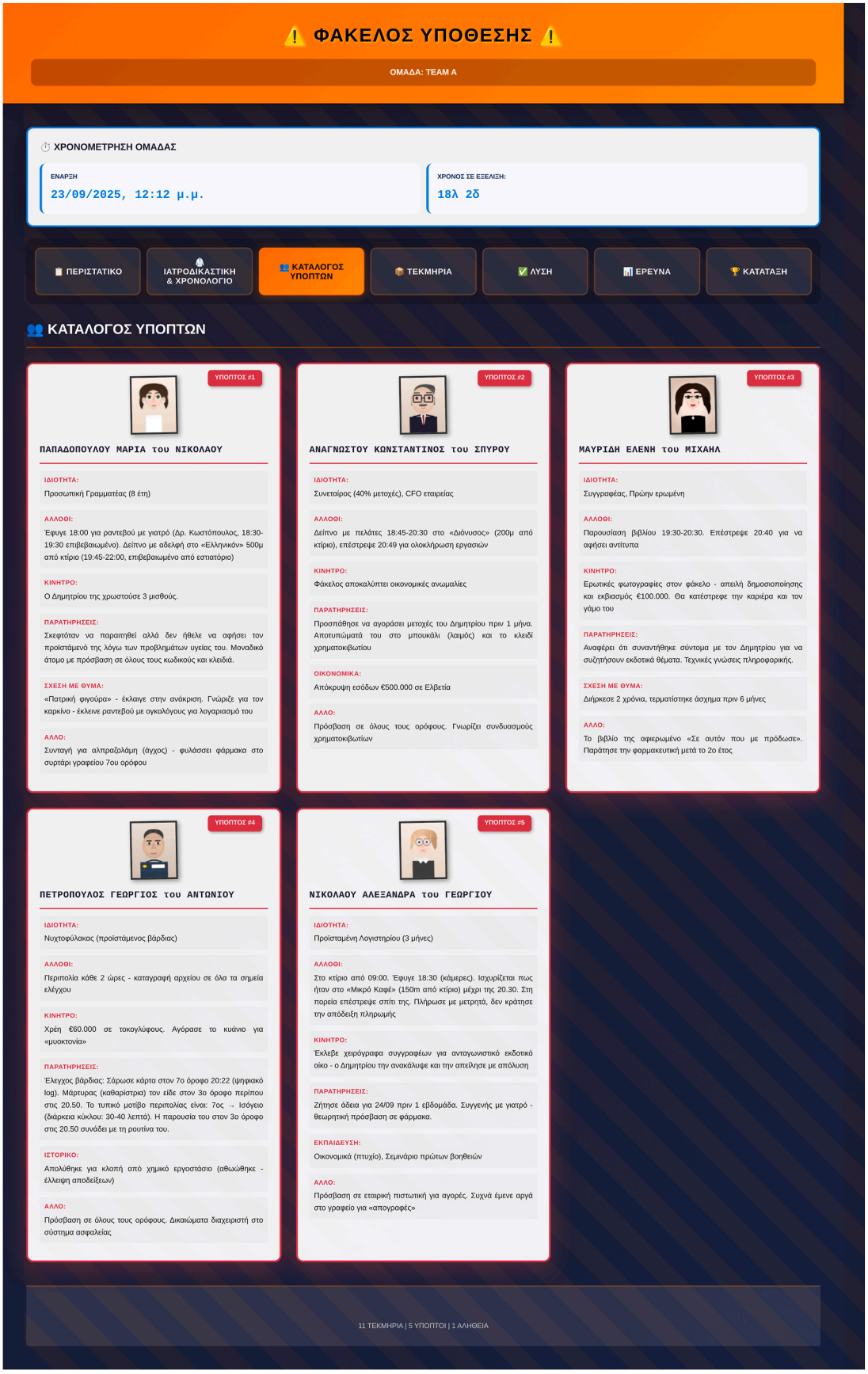

Fig. 3. Student-facing interface of the Mystery Crime Investigation environment (2/2).

Fig. 4. Augmented Reality evidence viewer in page context.

assessing beliefs about GenAI's utility for the mystery-solving task (e.g., "Using GenAI in this activity improved my problem-solving performance") whereas, PEU, was measured with six parallel items assessing effort expectations (e.g., "My interaction with GenAI was clear and understandable"). All items employed a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Internal consistency was good for PU ($\alpha = .79$), while PEU fell just below the conventional threshold ($\alpha = .69$).

A custom five-item scale was developed for the present study based on Bandura's (1997) guidelines for constructing self-efficacy scales (Appendix F). Specifically, each item was phrased as a task-specific confidence judgement of the form 'How confident are you that you can … '. Items of the 'Prompting Self-Efficacy Scale' (PSES) assessed confidence in formulating clear prompts, refining unsatisfactory responses, using AI for complex problems, critically evaluating AI outputs, and guiding AI through multi-step reasoning. Responses were recorded on a 7-point scale from 1 (*not at all confident*) to 7 (*completely confident*). Internal consistency was acceptable ($\alpha = .76$). Because the scale was newly developed, its structure was examined with an Exploratory Factor Analysis. Sampling adequacy was good (KMO = .81) and Bartlett's test of sphericity was significant ($\chi^2(10) = 177.47$, $p < .001$). Both the eigenvalue criterion and parallel analysis supported a single factor (first eigenvalue = 2.55, all others below .75) accounting for 39% of the variance, with loadings between .51 and .66 and corrected item-total correlations between .44 and .56 (Appendix F). Even so, the scale has no validation history in independent samples. As such, PSES is treated as an exploratory, study-specific, measure instead of a validated instrument and its associations with effectiveness are interpreted accordingly.

Prompts were categorised using the LBET framework (Fig. 5, Appendix G). Six categories were employed: *Searching* (locating facts and definitions), *Reading* (seeking explanations and summaries), *Interpreting* (drawing implications and connections), *Evaluating* (making judgments), *Organising* (structuring and synthesising information), and *Curating* (reflecting on AI use and validating reasoning). The categories were further aggregated into two stages: *E&A—comprising Searching, Reading, and Interpreting—and C&M—comprising* Evaluating, Organising, and Curating. Interpreting was treated as lower-order, in keeping with the placement of sense-making below evaluation and synthesis in the source taxonomies (Hmoud & Shaqour, 2024). In this task, interpreting evidence occasionally shaded into evaluating it; miscoding at that boundary would, if anything, render the E&A–C&M contrast conservative.

Fig. 5. The learning behaviour with emerging technologies framework.

3.5. Data analysis

Prompt coding was conducted by the lead researcher and verified through consensus coding with the co-authors (Saldaña, 2021). Inter-rater agreement, calculated on a stratified random subset of 94 prompts (20%), drawn proportionally across the 45 groups and independently coded by two researchers, resulted in a Cohen's κ of .87 which indicates excellent agreement (McHugh, 2012).

Hypothesis testing employed a combination of inferential techniques. For analyses involving individual-level measures (PU, PEU, PSES, experience, attitudes, and verification), scores were aggregated to group means, consistent with the group as the unit of analysis. The appropriateness of this aggregation was assessed using Intraclass Correlation Coefficients: ICC(1), indexing the proportion of variance in each measure attributable to group membership, and ICC(2), indexing the reliability of the group means. The dependent variable exists only at the group level, as each group receives a single effectiveness score. All analyses were, therefore, conducted between groups. The nesting of participants within groups matters for the individual-level predictors with the ICC indices being used to evaluate their aggregation to group means.

To examine development within the session, each group's ordered prompt sequence was divided into a first and a second half, with the middle prompt of odd-length sequences assigned to the second half. In greater detail, H1, which compared the prevalence of C&M versus E&A prompts, was tested with a paired-samples t-test at the group level. H2, which examined the relationship between the C&M ratio and effectiveness, was tested using Pearson correlation and linear regression. The regression followed a hierarchical specification, with prior experience and verification behaviour entered as additional predictors. Because the reasoning-quality component of the effectiveness score is rated by the same observers who logged the prompts, the relationship was re-estimated with effectiveness recomputed from its objective components alone (i.e., solution accuracy, completion time, and evidence interpreted); excluding the observer-rated 20-point reasoning component without rescaling. H3, which investigated non-linear experience effects, was tested using polynomial regression with linear and quadratic experience terms. H4, which examined the relationship between PU and effectiveness, was tested using correlation and regression analyses. H5, which examined verification behaviours, was tested using Spearman correlation given the ordinal nature of the verification frequency. Because teams were free to choose their GenAI tool, effectiveness and the C&M prompting ratio were additionally compared across the tools employed using one-way ANOVAs.

All analyses were conducted using SPSS (v.31) with the alpha level set at .05. Given the number of hypotheses, confirmatory and exploratory analyses were distinguished. H1 and H2 were treated as the confirmatory family and evaluated against a Bonferroni correction. H3, H4, and H5, together with all analyses involving PSES, were treated as exploratory and interpreted descriptively.

Qualitative data were collected through three open-ended questions administered post-activity (Appendix H). The questions asked participants to describe their group's AI interaction strategies, the challenges they encountered, and their reflections on using AI for collaborative tasks. Responses were analysed following Braun and Clarke's (2006) six-phase approach to thematic analysis which has been conducted inductively and at a semantic level. After familiarisation with the responses, two researchers independently generated initial codes across the full set of answers and then collated these codes into candidate themes which were reviewed against the coded extracts and the dataset as a whole before being defined and named. Coding was discussed iteratively until consensus was reached and disagreements were resolved by returning to the source responses. In keeping with a reflexive orientation, the researchers treated their familiarity with the LBET framework as a lens that could foreground cognitive-process themes and deliberately attended to responses that did not fit the dominant pattern. Therefore, divergent and contradictory views—such as difficulty trusting AI outputs and uncertainty about how to phrase prompts—were retained and are reported alongside the convergent themes. Exemplar quotes were selected to illustrate each theme on the basis of clarity and representativeness.

4. Results

4.1. Preliminary analyses

Prior to hypothesis testing, the distributions of key variables were examined for normality and outliers. Mystery effectiveness, the C&M ratio, and PU were approximately normally distributed at the group level, as confirmed by both visual inspection of histograms and the skewness and kurtosis values which were within conventional thresholds (skewness from −.63 to −.02, excess kurtosis from −.33 to .53) (Kim, 2013). No univariate outliers were detected (all z-scores $< |3.29|$). Parametric tests were therefore used. Between-group clustering (ICC(1)) was substantial for PSES (.26) but low for PU (.08, non-significant), PEU (.20), attitudes (≈.00), and, below zero, experience (−.02) and verification (−.04); the latter two indicating essentially no reliable between-group variance. The aggregated group means were correspondingly modest in reliability (ICC(2): PU = .24, PEU = .49, PSES = .57), so the group-level associations involving these measures are interpreted with caution. Likewise, the null associations for experience and verification should be read in light of this limited aggregate reliability. The tool comparisons showed no reliable differences in either effectiveness, $F(2, 42) = 1.81$, $p = .176$, or the C&M prompting ratio, $F(2, 42) = .11$, $p = .898$.

4.2. Descriptive statistics

Table 1 shows the descriptive statistics for the study variables at both individual and group levels. At the individual level ($N = 171$), participants reported moderately positive attitudes toward AI ($M = 3.28$, $SD = .35$ on a 5-point scale), high PU of GenAI ($M = 4.80$, $SD = .56$ on a 7-point scale), and high PEU ($M = 5.04$, $SD = .40$ on a 7-point scale). At the group level ($N = 45$), mean mystery effectiveness was 78.51 ($SD = 9.14$), an indicator of generally successful problem-solving performance. About two-thirds of groups (30 of 45) scored above the lowest effectiveness band (Table 4) thus, indicating that the task difficulty was appropriately calibrated.

4.3. Prompting behaviour distribution

Table 2 shows the distribution of the coded prompts ($N = 470$) across the LBET categories. 'Evaluating' prompts were most prevalent ($n = 114$; 24.3%), followed by 'Organising' ($n = 90$; 19.1%), 'Reading' ($n = 79$; 16.8%), 'Interpreting' ($n = 68$; 14.5%), 'Searching' ($n = 63$; 13.4%), and 'Curating' ($n = 56$; 11.9%). When aggregated by stage, C&M prompts constituted 55.3% of all prompts ($n = 260$), whereas E&A

Table 1
Descriptive statistics at individual and group levels.

Variable	*M*	*SD*	Range	*α*
Individual-level variables				
Age (years)	18.12	3.42	15–32	—
GAAIS	3.28	.35	2.13–4.38	.74
PU	4.80	.56	3.17–5.83	.79
PEU	5.04	.40	4.17–6.17	.69
PSES	3.95	.67	2.40–5.60	.76
Experience Score[a]	5.63	2.79	0–10	—
Group-level variables				
Mystery Effectiveness	78.51	9.14	55–96	—
Total Prompts	10.44	1.63	8–13	—
C&M Ratio	.55	.16	.13–.83	—

[a] A composite of duration (0–5) and frequency (0–5) of GenAI use.

Table 2
Distribution of prompts across the LBET categories.

Category	Stage	*n*	%	*M* per group
Searching	E&A	63	13.4	1.40
Reading	E&A	79	16.8	1.76
Interpreting	E&A	68	14.5	1.51
Evaluating	C&M	114	24.3	2.53
Organising	C&M	90	19.1	2.00
Curating	C&M	56	11.9	1.24
Total E&A	—	210	44.7	4.67
Total C&M	—	260	55.3	5.78

prompts constituted 44.7% (n = 210). The higher frequency of C&M prompts indicates that, on average, groups engaged more often in critical assessment and the structuring of information than in simple retrieval, although both prompt types were well represented.

4.4. Correlations

Table 3 presents the correlation matrix for group-level variables. Mystery effectiveness was strongly correlated with PU (r = .83, p < .001), PSES (r = .87, p < .001), PEU (r = .44, p = .003), and the C&M ratio (r = .40, p = .007). Non-significant correlations were found between effectiveness and experience score (r = .14, p = .346), verification behaviour (r = .13, p = .40), and attitudes toward AI (r = .03, p = .865). PU and PSES were strongly intercorrelated (r = .81, p < .001). Experience correlated moderately with PEU (r = .58, p < .001) but only weakly with PU (r = .23, p = .121), thus indicating that experience relates to perceived ease but not to perceived value.

4.5. Hypothesis testing

H1 predicted that groups would generate more higher-order (C&M) prompts than lower-order (E&A) prompts. A paired-samples t-test compared C&M and E&A counts within groups. Groups produced a mean of 5.78 C&M prompts (SD = 1.98) against 4.67 E&A prompts (SD = 1.52), a difference of 1.11 prompts per group. The difference was significant (t(44) = 2.38, p = .022) and also carried a small-to-medium effect (d = .36). The same prompts, divided at the midpoint of each session, show where that surplus accumulated. C&M prompts made up 40.2% of the prompts issued in the first half (SD = 27.8%) and 68.2% in the second (SD = 25.3%), a within-session increase that was itself significant (t(44) = −4.39, p < .001, d = .66) and consistent with groups moving from information gathering towards critical synthesis as their understanding of the mystery developed. With a mean of 10.44 prompts per group (Table 1), each half rests on roughly five prompts, so the temporal contrast is read here as a coarse, exploratory indication of within-session change. The confirmatory comparison stands independently of it and as such the first hypothesis can be confirmed: groups produced more higher-order than lower-order prompts when collaborating with GenAI.

H2 predicted that a higher proportion of C&M prompts would be associated with greater problem-solving effectiveness. To test the hypothesis, a Pearson correlation analysis was conducted. The relationship between the C&M ratio and mystery effectiveness was positive and significant, r = .40, p = .007 (Fig. 6).

Table 4
Prompting quality and technology acceptance across effectiveness bands.

Effectiveness Band	n	Effectiveness[a]	C&M Ratio	PU Mean
Q1 (Low: 55–74)	15	68.40	.48	4.47
Q2 (75–79)	8	77.13	.49	4.78
Q3 (80–85)	11	82.36	.56	4.91
Q4 (High: 86–96)	11	89.45	.67	5.15

[a] Scores are based on group performance.

In the equivalent regression, the C&M ratio accounted for 16% of the variance in effectiveness, B = 23.33, SE = 8.23, t(43) = 2.83, p = .007, R^2 = .16. Across effectiveness bands (Table 4), the mean C&M ratio was .67 in the highest band and .48 in the lowest. The association was stable under alternative compositions of the effectiveness score. With the four components weighted equally, the correlation was r = .42, p = .004; with an evidence-emphasised weighting (solution accuracy 30%, completion time 10%, evidence interpretation 40%, reasoning quality 20%), it was r = .37, p = .012. The alternative composites correlated at r ≥ .96 with the original score. When effectiveness was recomputed from its objective components alone (Section 3.5), the association remained positive and significant, r = .38, p = .011, so it is not an artefact of shared reasoning content. In addition, total prompt volume was entered as a covariate and the C&M coefficient remained significant, B = 17.60, p = .030; the relationship is therefore not attributable to more active groups producing more prompts of every kind. The primary H2 association remained significant after Bonferroni correction; the covariate-adjusted coefficient was significant at the conventional level, though it does not clear the corrected threshold (α = .025). On these grounds the second hypothesis can be confirmed: groups that directed a higher proportion of higher-order prompts to the system solved the mystery more effectively.

H3 predicted an inverted-U relationship between prior GenAI experience and problem-solving effectiveness, with moderate experience returning the strongest outcomes. To test this hypothesis, a Polynomial Regression analysis was conducted. The linear model was not significant, B = .95, p = .346, R^2 = .02 and the addition of the quadratic term did not significantly improve the model fit either, ΔR^2 = .007, F-change (1, 42) = .31, p = .579. It is, therefore, concluded that the third hypothesis should be rejected: prior experience was unrelated to effectiveness in this collaborative context, in neither its linear nor its curvilinear form.

H4 predicted that higher PU would be associated with greater problem-solving effectiveness. A strong positive correlation (Fig. 7) was found between PU and mystery effectiveness, r = .83, p < .001, R^2 = .69. The corresponding regression coefficient, B = 23.70 (SE = 2.45), t (43) = 9.67, p < .001, amounts to approximately 24 additional points of effectiveness for each one-point increase in PU on the 7-point scale. It should be noted that PU was administered immediately after the task, with items referencing the task itself, so part of the association may

Table 3
Correlation matrix for group-level variables.

Variable	1	2	3	4	5	6	7	8
1. Effectiveness	—							
2. C&M Ratio	.40**	—						
3. PU	.83***	.26	—					
4. PEU	.44**	.11	.34*	—				
5. PSES	.87***	.25	.81***	.46**	—			
6. Experience	.14	.02	.23	.58***	.33*	—		
7. Verification	.13	.02	.09	.17	.05	−.00	—	
8. GAAIS	.03	.35*	−.02	.11	.04	.02	.03	—

*p < .05, **p < .01, ***p < .001.

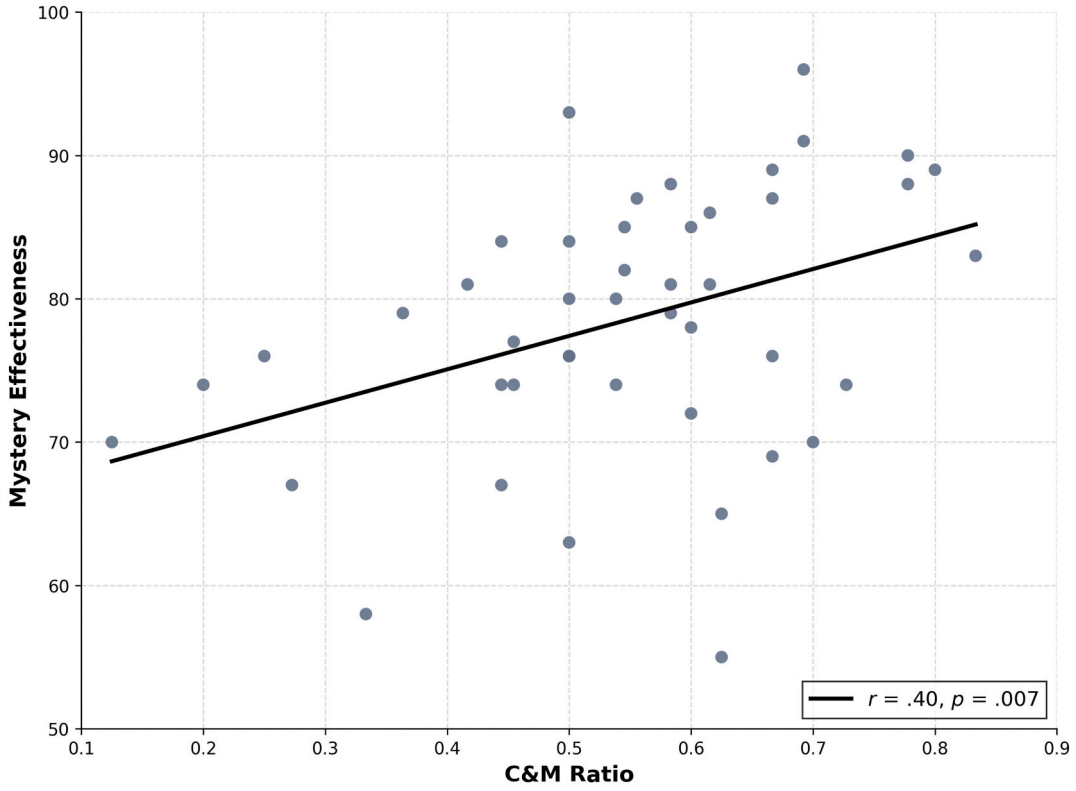

Fig. 6. Relationship between C&M ratio and mystery effectiveness scores across groups.

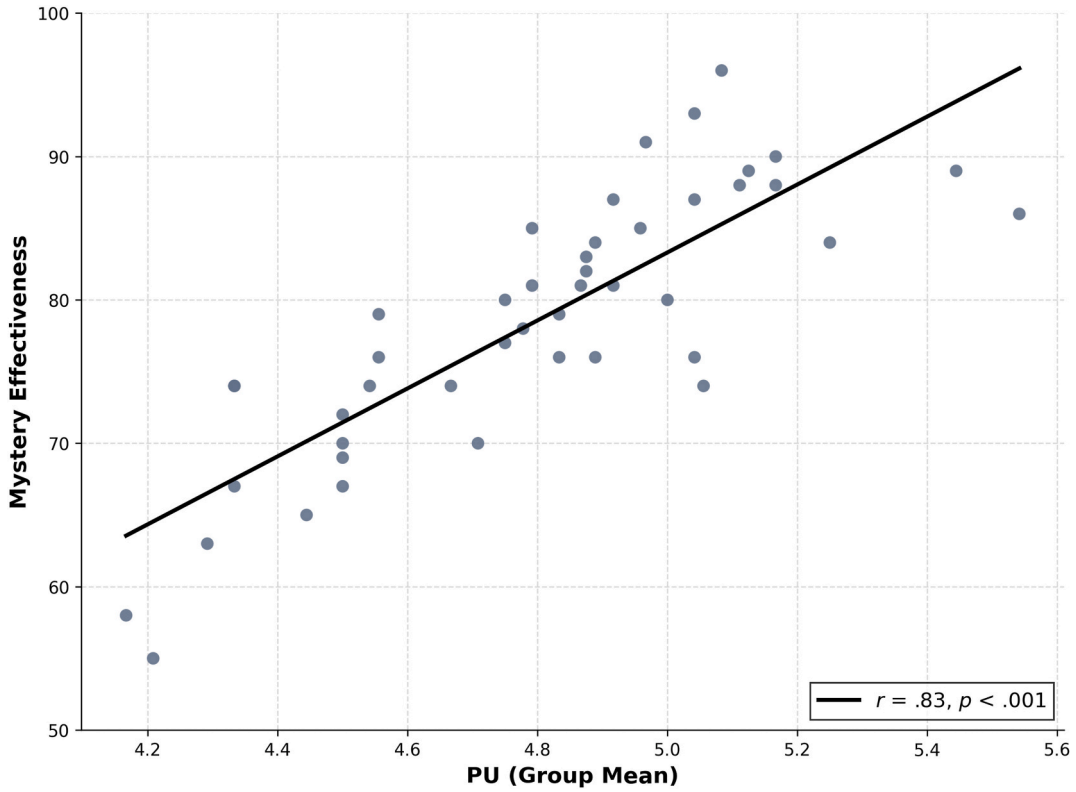

Fig. 7. Relationship between PU and mystery effectiveness at group level.

reflect shared task-outcome content. The fourth hypothesis can be confirmed in a concurrent sense only: group-level PU and effectiveness were strongly related, though the correlation does not by itself establish that usefulness beliefs produced the performance.

H5 predicted that more frequent verification of the AI-generated responses would be associated with greater effectiveness. The

correlation was positive but not significant, ρ = .16, p = .294 (Pearson r = .13; Table 3). It is concluded that the fifth hypothesis should be rejected: how often groups checked the system's output was unrelated to how effectively they solved the mystery.

4.6. Regression analysis

To examine the unique contribution of each predictor, a Hierarchical Regression Analysis was conducted with mystery effectiveness as the dependent variable (Table 5).

In Step 1, the C&M ratio alone predicted effectiveness, $R^2 = .16$, F(1, 43) = 8.03, p = .007. Adding prior experience in Step 2 did not improve the model, $\Delta R^2 = .02$, F-change(1, 42) = .92, p = .343, with the C&M coefficient remaining essentially unchanged (B = 23.14, SE = 8.24, t = 2.81, p = .008). Adding verification behaviour in Step 3 again contributed no incremental variance, $\Delta R^2 = .01$, F-change(1, 41) = .71, p = .406. In the final model the C&M ratio remained the only reliable predictor (B = 22.97, SE = 8.27, t = 2.78, p = .008), with neither prior experience (B = .89, p = .343) nor verification (B = 2.21, p = .406) approaching significance; the three predictors together accounted for 19% of the variance in effectiveness, F(3, 41) = 3.19, p = .033.

4.7. Qualitative evaluations

Open-ended responses (n = 22 of the 171 participants, 12.9%) offered further insight into the collaborative experience with GenAI. The analysis revealed three strategy-related themes: (a) exploratory questioning, (b) structured information gathering, and (c) higher-order synthesis; two challenge-related themes: (a) trust calibration and (b) prompt formulation difficulties; and three metacognitive insights: (a) specificity in questioning, (b) verification necessity, and (c) critical thinking alongside AI.

Regarding prompting strategies, many participants described a progression from exploratory questioning toward more structured approaches, moving from broad requests for explanation to targeted questions about individual suspects and, later, to requests for synthesis such as organising the evidence and identifying contradictions. Perceived challenges clustered around trust calibration and query construction; participants found it hard to judge which AI responses to trust and how to phrase questions productively. The lessons learned centred on metacognitive awareness, covering specificity in questioning, the need to verify AI answers, and the view that the tool does not replace the group's own thinking. Given the modest response rate, the themes are presented as an illustrative complement to the quantitative findings; Table 6 reports each theme with a representative excerpt.

5. Discussion

The aim of the current study was to examine what characterises effective collaborative problem-solving when groups of students work with GenAI. The findings showed that groups engaged more often in higher-order than lower-order prompting, that the proportion of higher-order prompts was associated with task effectiveness, and that PU and PSES were strongly correlated with it. In contrast, prior experience and verification behaviour showed no reliable association with effectiveness.

Table 5
Multiple regression analysis of group effectiveness.

Predictor	B	SE	t	p	R^2	ΔR^2
Step 1					.157	.157**
C&M Ratio	23.33	8.23	2.83	.007		
Step 2					.175	.018
C&M Ratio	23.14	8.24	2.81	.008		
Experience	.89	.93	.96	.343		
Step 3					.189	.014
C&M Ratio	22.97	8.27	2.78	.008		
Experience	.89	.93	.96	.343		
Verification	2.21	2.63	.84	.406		

**p < .01.

Table 6
Themes from the open-ended responses with representative excerpts.

Theme cluster	Theme	Representative excerpt
Strategies	Exploratory questioning	"We first asked the AI to explain the case, then asked specific questions about each suspect"
Strategies	Structured information gathering	"We used the AI like a search engine at first, then started asking it to make connections"
Strategies	Higher-order synthesis	"We asked the AI to organise the evidence and identify contradictions"
Challenges	Trust calibration	"It was hard to know which AI responses to trust"
Challenges	Prompt formulation difficulties	"We didn't know how to phrase questions to get good answers"
Metacognitive insights	Specificity in questioning	"Being specific in your questions gives better results"
Metacognitive insights	Verification necessity	"You need to verify AI answers, not just accept them"
Metacognitive insights	Critical thinking alongside AI	"AI is a tool, not a solution—we still need to think"

5.1. Higher-order prompting in collaborative GenAI use

The greater prevalence of C&M than E&A prompts aligns with recent evidence on GenAI's cognitive effects with Deng et al. (2025) reporting positive effects of ChatGPT on academic performance, motivation, and higher-order thinking. As such, the view that GenAI interaction can foster cognitive engagement, when appropriately structured, is supported. Heung and Chiu's (2025) meta-analysis similarly found that ChatGPT interventions enhanced student engagement through active cognitive processing which further supports the argument that GenAI can scaffold higher-order thinking.

The finding that C&M prompt proportion was positively associated with task effectiveness resonates with Crawford et al.'s (2023) study which demonstrates that higher-level learning occurs when students use GenAI to construct and augment knowledge through a mastery approach instead of passive consumption. Groups in our study that invested more interaction effort in consolidating activities (i.e., requesting synthesis, evaluation, and strategic guidance) achieved outcomes consistent with what Crawford et al. (2023) term as "knowledge-building orientation". Essel et al. (2024), who examined university students' ChatGPT interactions, reported also positive impact on critical, reflective, and creative thinking; findings which align with our observation that cognitively demanding prompts returned superior outcomes.

One explanation for the C&M-effectiveness relationship lies in the cognitive demands of higher-order prompting. Essel et al. (2024) found that ChatGPT use reduced cognitive load, thereby suggesting that AI can free cognitive resources for analysing and solving complex problems when users engage strategically. Our findings extend the aforementioned observation; groups using C&M prompts may have offloaded routine information gathering to GenAI while directing their own cognitive resources toward integration and evaluation—the division of labour that Järvelä et al. (2025) characterise as "successful hybrid intelligence".

The task structure in our study may have scaffolded the adaptive prompting pattern, which is also consistent with Wu et al.'s (2026) finding that ChatGPT's effects on learning outcomes are moderated by subject (task domain). The temporal pattern observed in our study provides further support for this interpretation. Groups shifted from E&A prompts in the first half of sessions to C&M in the second half, thus suggesting that higher-order prompting emerged as groups developed

task understanding instead of being an initial default. The progression mirrors Gonsalves's (2026) finding that effective AI-assisted learning requires recursive engagement across cognitive domains with learners who progress from exploratory to consolidating behaviours achieving stronger outcomes.

5.2. Technology acceptance beliefs and task performance

Evidence that PU carries beyond intentions to outcomes is accruing. Boubker (2024) found that the PU of ChatGPT raised use and satisfaction, which in turn enhanced students' learning outcomes. The practical reading is that demonstrating a tool's concrete utility within meaningful tasks may matter more for productive engagement than promoting adoption as such. The finding that PU was associated with task effectiveness points beyond the adoption intentions that dominate TAM research toward performance-relevant engagement in collaborative GenAI use. Our result is in line with the conclusions that both Dahri et al. (2024) and Budhathoki et al. (2024) reached wherein PU is identified as the strongest attitudinal predictor. The direction of this association remains open, since users who believe GenAI enhances performance may engage more strategically, whereas more effective groups may equally come to regard the tool as more useful. Given that the PU items referenced the task itself and were completed immediately after it, the strength of the correlation should also be read with this measurement proximity in mind.

5.3. Prompting self-efficacy and collaborative effectiveness

The finding that PSES— an exploratory measure—showed the strongest correlation with task effectiveness should be interpreted with caution in view of the following considerations that frame the result. First, PSES is a study-specific instrument; although its unidimensionality and internal consistency proved satisfactory, it has no validation history, so its convergent and discriminant properties are unknown. Second, the measure was completed after the task. As such, any confidence judgements made in the immediate aftermath of success or failure tend to absorb outcome information. Third, the association is consistent with emerging evidence that self-efficacy beliefs relate to GenAI engagement (Shahzad et al., 2025). An association which is potentially bidirectional since confidence may support more effective prompting and succeeding at the task may equally raise confidence. Liang et al. (2023) reported, in the same direction, that the relationship between students' GenAI interaction and achievement was mediated serially by self-efficacy and cognitive engagement. Therefore, establishing whether prompting self-efficacy is an antecedent, a consequence, or both will require designs in which confidence is measured before the collaborative task.

5.4. Experience-performance effects

The null association between prior experience and effectiveness is puzzling, given the theoretical grounding in automation complacency and the mixed empirical evidence on AI experience effects. As we did not anticipate this pattern, the accounts that follow are post-hoc, and each warrants direct testing. Recent HCI evidence suggests that prompting behaviour is shaped by user dispositions and by the task at hand (Wyszynski et al., 2026). It also indicates that users do not readily develop systematic prompting strategies through use alone (Zamfirescu-Pereira et al., 2023). Time with the technology may therefore accumulate exposure without accumulating strategy. Experience may also carry a cost that offsets its benefit, since reliance on automated aids draws attention away from monitoring and verification (Parasuraman & Manzey, 2010). The likeliest account, though, concerns the measure itself. AI literacy comprises distinguishable dimensions (Ng et al., 2021) and general exposure to GenAI may not translate into the domain-specific collaborative problem-solving competence our task required.

5.5. Verification behaviour and task performance

The finding that verification behaviour did not predict task effectiveness was also unexpected, given recent emphasis on critical evaluation of AI outputs. Dwivedi (2025) developed the 'ChatGPT Fact-Check' assignment, to teach students evidence-based reasoning by evaluating AI-generated claims, and found that structured verification activities enhanced critical thinking. The accounts that follow are again post-hoc. One is measurement: our single-item measure may have captured intentions more than actions, weakening any true relationship with outcomes. Another is that verification matters most when the AI outputs contain errors. In their review of ChatGPT, Bettayeb et al. (2024) noted that the technology provides generally accurate information for common queries, with verification becoming critical for novel or complex claims. If the system's mystery-solving suggestions were largely accurate, verification effort may have represented unnecessary cognitive expenditure. A third account comes from the collaborative context itself, which may have supplied implicit verification through group discussion: members debating the system's suggestions may have evaluated the outputs collectively, with no individual reporting a high verification tendency. The reading aligns with Atchley et al.'s (2024) model of Human-AI collaboration in Higher Education which positions metacognitive monitoring as a distributed function in collaborative settings.

More broadly, the null findings for H3 (non-linear experience) and H5 (verification) raise questions about the applicability of ACT to conversational GenAI. The framework of Parasuraman and Manzey (2010) was developed primarily in contexts involving automated decision aids, that is, systems that provide discrete recommendations which users can accept or reject. Conversational GenAI differs in that it engages users in open-ended dialogue, produces varied outputs depending on prompt formulation, and requires active participation to generate useful responses. Such characteristics may attenuate complacency effects, with the dialogic nature of the interaction providing inherent safeguards against the passive acceptance that characterises automation complacency in other domains.

5.6. Experience, Perceived Usefulness, and task effectiveness

A descriptive pattern among experience, technology-acceptance beliefs, and effectiveness merits comment. Prior experience was related to PEU but showed only a weak, non-significant association with PU. In addition, neither belief lay on a demonstrable path from experience to effectiveness. The asymmetry suggests that GenAI experience brings fluency (i.e., the sense that the tool is easy to use) without a corresponding appreciation of its genuine utility.

The finding that GAAIS showed no relationship with effectiveness further supports the specificity of beliefs that matter. General attitudes were nonetheless modestly associated with the proportion of higher-order prompting even though they were unrelated to effectiveness itself, thus suggesting that positive dispositions toward AI may shape how groups engage the tool without translating into better outcomes. Chiu et al.'s (2023) review of AI in education emphasised that AI literacy involves specific competencies instead of general dispositions. Broad positive attitudes toward AI appear insufficient—what matters is situation-specific beliefs about AI's utility for particular tasks; beliefs that may develop through the quality of experience as opposed to its frequency. Cotton et al. (2024) similarly argued that academic integrity considerations require layered understanding of AI capabilities that general attitudes fail to capture.

6. Implications

The study has implications for both theoretical development and practice.

First, the association between C&M prompting and effectiveness is

correlational and does not, by itself, establish that training learners to produce higher-order prompts will improve performance. What it does license is a targeting claim: higher-order cognitive behaviours are the observable markers that accompany effective collaboration and they—more than prompt formatting or keyword optimisation—are, therefore, reasonable candidates for AI literacy curricula to address, pending experimental confirmation that cultivating them returns gains. The implication aligns with the AI literacy framework of Ng et al. (2021) which positions evaluation and creation among its core competencies. Similarly, Ansari et al.'s (2024) review concluded that educators should design assessment tasks requiring critical thinking and human intelligence when working with AI. In this regard, our findings are consistent with curricula that scaffold the progression from exploratory to consolidating AI interactions.

A second implication concerns PU in AI engagement. Instead of assuming that exposure develops competence, educators might focus on demonstrating concrete utility through meaningful tasks. The recommendation resonates with Chiu's (2024b) framework for GenAI integration which emphasises authentic applications where AI adds genuine value. Chiu et al. (2023) also found that AI assistance can either support or undermine student agency—depending on how tasks position AI's role. The concurrent PU–effectiveness association observed here is consistent with tasks that reveal AI's genuine utility building the motivational orientations that sustain effective collaboration. Chiu's (2024a) self-regulated learning framework for AI-assisted writing similarly emphasises reflection on AI's contribution as essential for developing productive use patterns.

From a theoretical perspective, the findings connect technology-acceptance beliefs to performance-relevant behaviour in collaborative AI contexts. The LBET framework proved workable for capturing variation in prompting quality, thus supporting its application to GenAI interaction research by illustrating the value of Bloom-based taxonomies (Hmoud & Shaqour, 2024; Hui, 2025) for categorising human-AI interaction in collaborative settings. The inconclusive findings for experience and verification suggest that existing theoretical models require adaptation for Human-GenAI collaboration where open-ended conversational interaction differs from the automated decision aids that inform traditional automation research (Lebiere et al., 2021).

7. Limitations and future directions

The study's limitations provide directions for future research.

First, because outcomes were measured at a single point, the study cannot reveal whether the observed associations persist. Bastani et al. (2025) demonstrated that GenAI learning effects can be ephemeral, appearing in immediate assessment but disappearing when support is withdrawn. Longitudinal designs tracking whether prompting quality and PU predict sustained learning, beyond specific AI-assisted sessions, would address this limitation.

Second, the single task type may not generalise to other collaborative AI applications. Wu et al. (2026) found that ChatGPT's effectiveness varies by discipline, intervention duration, and instructional mode, while educational level and knowledge type did not significantly moderate outcomes. Replication across creative production, technical problem-solving, and decision-making tasks would clarify boundary conditions for our findings. Testing whether automation-complacency predictions hold in more constrained GenAI applications—such as code completion or AI-generated summaries, where the interaction more closely resembles traditional automation—would further delimit the theory's scope. In view of that, future research should also examine whether collaborative AI effects differ from individual AI-assisted learning, as most existing meta-analyses aggregate across collaborative and individual AI use (Heung & Chiu, 2025).

Third, reliance on self-reported measures for experience and verification introduces potential response biases. The verification measure, in particular, consisted of a single ordinal item, whose unknown reliability attenuates any true association with effectiveness. Recent studies (Crawford et al., 2023; Steiss et al., 2024) have moved toward behavioural trace data to analyse reflections with emerging tools by tracking actual conversation logs. Future research might incorporate logged prompts, response times, and external verification actions to complement self-report assessment. Future work could also separate passive AI consumption from active skill-building engagement and task-aligned from task-misaligned experience to capture effects that a single composite experience measure may obscure.

Fourth, the reasoning-quality component of the effectiveness score was rated by a single GM per group, so inter-rater reliability could not be established for that component. The effectiveness composite also weighted solution accuracy most heavily, with completion time, evidence interpreted, and reasoning quality each contributing 20 points. Because the time component rewards faster completion, it may partly conflate efficient problem-solving with reduced deliberation, although the higher-order-prompting association persisted when effectiveness was recomputed from its objective components alone.

Fifth, the open-ended qualitative responses were provided by a small fraction of the participants' pool ($n = 22$), thus limiting the representativeness of the themes identified. As such, the qualitative findings should be read as illustrative of the quantitative patterns.

Sixth, the competitive framing—although confined to the between-group level and expected on theoretical grounds to support engagement (Landers et al., 2017; Sailer & Homner, 2020)—may have shifted how groups balanced speed against deliberation; its influence cannot be separated from that of the task itself.

Finally, the student sample may not represent AI collaboration patterns in professional contexts where stakes and expertise levels differ.

8. Conclusion

In the present study, we examined how the prompting behaviour of groups developed over a collaborative mystery-solving task with GenAI and how it related to the effectiveness of the collaboration. Groups produced a modest majority of higher-order prompts, shifted from exploratory toward evaluative and synthetic prompting as the sessions advanced, and performed better the larger the share of higher-order prompts they produced; the association withstood controls for prompt volume and alternative compositions of the effectiveness score. Prior experience, general attitudes, and self-reported verification showed no reliable association with effectiveness. PU and PSES, in contrast, covaried strongly with it, as concurrent correlates measured after the task. It is concluded that what groups do with GenAI is more informative about the effectiveness of the collaboration than who their members are in terms of prior exposure or general disposition. For practice, the results direct attention toward cultivating evaluative, organisational, and synthetic engagement with AI. For research, they argue for group-level, time-resolved designs and for experimental tests of whether higher-order prompting can be taught to advantage.

CRediT authorship contribution statement

Athanasios Christopoulos: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft. **Stylianos Mystakidis:** Conceptualization, Methodology, Validation, Writing – review & editing. **Isidoros Perikos:** Validation, Writing – review & editing. **Aikaterini Bourazeri:** Data curation, Validation, Writing – review & editing. **Andria Michael:** Writing – review & editing. **Mikko-Jussi Laakso:** Funding acquisition, Supervision, Writing – review & editing.

Ethics declaration

The study was performed in compliance with the laws, regulatory frameworks and guidelines applicable where the research took place,

and in accordance with the ethical standards of the Declaration of Helsinki (1964) and its later amendments. Ethics committee approval was not required under the relevant laws and the national research-ethics guidelines applicable to the coordinating institution. Under those guidelines, an ethical review statement is mandatory only for research that deviates from the principle of informed consent, intervenes in participants' physical integrity, focuses on minors under 15 years of age without a guardian's consent, exposes participants to exceptionally strong stimuli, risks mental harm beyond that of normal daily life, or could threaten the safety of participants or researchers. The present non-interventional, observational study met none of these conditions. All data were de-identified, and confidentiality and data protection were maintained throughout the study and in all reporting of the results.

Declaration on informed consent

Written informed consent to take part in the study and to publish the results was obtained from all participants prior to participation. For participants under 18 years of age, written informed consent was obtained from a parent or legal guardian, and assent was obtained from the minor. All participants were informed of the purpose of the study, the procedures involved, and their rights as participants, including the right to privacy and confidentiality. Participation was voluntary, and participants were free to withdraw at any time without consequence.

Funding

The study is part of the EDUCA Flagship project, funded by the Research Council of Finland [grant numbers 358924, 358947].

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank all participating students and volunteers who contributed their time to support the study.

APPENDIX A. Game Master Observation Rubric

During each session, the group's assigned Game Master completed a structured observation rubric. Each criterion was recorded live, together with free-text general notes.

Observation domain	Criteria recorded
AI use patterns	Frequency of queries to the AI; prompt quality; verification of AI answers
Team collaboration	Active discussion among members; exchange of information and evidence; division of tasks
Problem-solving approach	Systematic analysis of the evidence; cross-referencing of evidence pieces; critical questioning of conclusions
Engagement and motivation	Enthusiasm and energy; persistence despite difficulties; focus and attention
General notes	Free-text observations recorded during the session

Reasoning Quality Scoring Rubric

The reasoning-quality component (*R*) was rated by the Game Master who was overseeing the group's problem-solving approach and supporting argument, using the predefined observation rubric—which included: systematic analysis of the evidence, cross-referencing of evidence pieces, and critical questioning of conclusions—on a continuous 0–1 scale capturing logical coherence and evidence-based reasoning. Intermediate values (for example, .60) were permitted where a group's reasoning fell between two anchors.

Score	Anchor description
.00	No coherent reasoning. The conclusion is asserted without logical justification; available evidence is ignored, misread, or contradicted.
.25	Minimal reasoning. A conclusion is offered, but the logical chain is fragmentary or internally inconsistent; only isolated pieces of evidence are cited, and key inconsistencies are left unaddressed.
.50	Partial reasoning. A broadly logical argument is present but incomplete; some relevant evidence is integrated, but noticeable gaps, unsupported leaps, or unexamined alternative explanations remain.
.75	Sound reasoning. A coherent, well-sequenced argument that integrates most of the relevant evidence and addresses the main competing explanation, with only minor gaps or under-justified steps.
1.00	Fully coherent reasoning. A complete, internally consistent chain of inference that integrates the available evidence, explicitly rules out the principal alternative explanations (e.g., other suspects), and justifies the final conclusion.

Outcome Measure – Mystery Effectiveness

The observer-rated group performance measure (i.e., the Mystery Effectiveness variable) is calculated on a continuous scale (0-100) with all participants in a group sharing the same score.

Criterion	Description	Weight
Solution Accuracy	Correctness of the final mystery solution; identification of key elements	40%
Completion Time	Time taken to reach a solution, relative to the 90-min cap	20%
Reasoning Quality	Logical coherence and evidence-based argumentation in solution	20%
Evidence Interpretation	Number of evidence pieces correctly interpreted (0–11)	20%

APPENDIX B

Demographics Questionnaire

Administration: Pre-activity.

Item	Variable	Question	Response Options
1.1	Gender	Please state your gender.	Male; Female; Prefer not to say; Other
1.2	Age	Please state your age.	Open numeric response
1.3	Education Level	Please state your educational level.	Secondary Education; Undergraduate; Postgraduate
1.4	Educational Field	Please state your field of study.	Science and Technological Sciences; Economics and Information Technology; Humanities; Health and Life Sciences; Computer Science; Engineering; Other [Please specify]

APPENDIX C

GenAI Experience Questionnaire

Branched design based on prior GenAI use. Administration: Pre-activity.

Branching Question

Item	Question	Response Options
2.0	Have you ever used Generative AI tools (e.g., ChatGPT, Claude, Gemini, Copilot)?	Yes → Branch A; No → Branch B

Branch A: For GenAI Users

Item	Question	Response Options
2.A1	How long have you been using GenAI tools?	<1 month; 1–3 months; 4–6 months; 7–12 months; >1 year
2.A2	How often do you use GenAI tools?	Daily; 4–6×/week; 2–3×/week; Once/week; 2–3×/month
2.A3	Which GenAI tools do you use?	ChatGPT; Claude; Google Gemini; Microsoft Copilot; Other (specify)
2.A4	For what purposes do you typically use GenAI?	Homework; Study assistance; Research; Writing; Creative projects; Other (specify)
2.A5	How would you rate your prompting capability?	1 = Not at all capable – 5 = Extremely capable
2.A6	How much do you trust GenAI responses?	1 = Not at all – 5 = Completely
2.A7	How often do you verify GenAI responses?	Always; Often; Sometimes; Rarely; Never
2.A8	What is your biggest challenge when using GenAI?	Formulating effective questions; Doubt about accuracy; Technical issues; No challenges; Other

Branch B: For Non-Users

Item	Question	Response Options
2.B1	Why haven't you used GenAI tools?	Prefer traditional methods; Do not need them; Do not trust AI; Lack access; Other
2.B2	How aware are you of what GenAI can do?	1 = Not at all aware – 5 = Fully informed
2.B3	What concerns do you have about GenAI?	Academic integrity; Accuracy; Privacy; No concerns; Other
2.B4	How likely are you to use GenAI in the future?	Very unlikely; Unlikely; Neutral; Likely; Very likely
2.B5	What would motivate you to try GenAI?	Training; Free access; Teacher recommendation; Peer influence; Other

Common Questions (All Participants)

Item	Question	Response Options
2.O1	How important do you think GenAI will be for your future career?	1 = Not at all important – 5 = Extremely important
2.O2	How adequate is the GenAI education you have received?	1 = Extremely inadequate – 5 = Extremely adequate

APPENDIX D

General Attitudes towards Artificial Intelligence Scale (GAAIS)

Administration: Pre-activity. Scale: 1 = Strongly Disagree to 5 = Strongly Agree.

Positive Attitudes Subscale

Item	Statement
GAAIS_P1	I am impressed by what Artificial Intelligence can do.
GAAIS_P2	I am interested in using Artificial Intelligence in my daily life.
GAAIS_P3	I think Artificial Intelligence is exciting.
GAAIS_P4	I think Artificial Intelligence is beneficial for society.
GAAIS_P5	Artificial Intelligence can perform better than humans in some tasks.

Negative Attitudes Subscale (Reverse Scored)

Item	Statement
GAAIS_N6	I think Artificial Intelligence is likely to make mistakes.
GAAIS_N7	I think Artificial Intelligence is dangerous.
GAAIS_N8	I feel uncomfortable around Artificial Intelligence systems.

APPENDIX E

Technology Acceptance Model Scales

Administration: Post-activity. Scale: 1 = Strongly Disagree to 7 = Strongly Agree.

Perceived Usefulness

Item	Statement
PU1	Using GenAI in this activity improved my problem-solving performance.
PU2	Using GenAI increased my productivity in the task.
PU3	Using GenAI enhanced my effectiveness in the activity.
PU4	I found GenAI useful for completing this activity.
PU5	GenAI made it easier to accomplish the task requirements.
PU6	Overall, I found GenAI to be advantageous in this activity.

Perceived Ease of Use

Item	Statement
PEU1	Learning to use GenAI for this activity was easy.
PEU2	I found it easy to get GenAI to do what I wanted.
PEU3	My interaction with GenAI was clear and understandable.
PEU4	I found GenAI flexible to interact with.
PEU5	It was easy to become skillful at using GenAI.
PEU6	Overall, I found GenAI easy to use.

APPENDIX F

Prompting Self-Efficacy Scale (PSES)

Administration: Post-activity.
Custom instrument. Scale: 1 = Not at all confident to 7 = Completely confident. Score Range: 1.0–7.0 (mean).
Instructions: Please rate how confident you feel about each of the following abilities based on your experience in today's activity.

Item	How confident are you that you can …
PSES_1	… formulate clear and effective prompts to get useful responses from AI?
PSES_2	… refine and improve your prompts when AI responses are not satisfactory?
PSES_3	… use AI effectively to solve complex problems you couldn't solve alone?
PSES_4	… critically evaluate AI responses for accuracy and usefulness?
PSES_5	… guide AI through multi-step reasoning to reach a solution?

Psychometric Properties

The scale was constructed for the present study following Bandura's (1997) guidelines. The Factor Analysis supporting its unidimensional structure is reported in the Instruments section of the main text; item-level statistics are reported below.

Item	*M*	*SD*	Factor loading	Corrected item-total *r*	α if item deleted
PSES_1	3.99	.93	.65	.54	.71
PSES_2	3.85	.98	.66	.56	.70
PSES_3	3.92	.95	.66	.56	.70
PSES_4	3.92	.90	.63	.53	.71
PSES_5	4.07	.93	.51	.44	.75

APPENDIX G

Learning Behaviour with Emerging Technologies Framework

Adapted from Bloom's Digital Taxonomy. Application: Researcher coding of prompt logs.

Category	Code	Stage	Description
Searching	SEA	E&A	Basic information retrieval; asking for definitions, facts, or descriptions
Reading	REA	E&A	Comprehension-focused; asking for explanations, summaries, or clarifications
Interpreting	INT	E&A	Application of information; asking for examples, implications, or connections
Evaluating	EVL	C&M	Critical assessment; asking AI to compare, judge, critique, or assess information
Organising	ORG	C&M	Synthesis and structuring; asking AI to organise, prioritise, or create plans
Curating	CUR	C&M	Metacognitive oversight; reflecting on AI use, validating reasoning, refining approach

APPENDIX H

Open-Ended Questions

Administration: Post-activity.

Item	Question
OE_1	What strategies did your group use when interacting with AI to solve the mystery? Please describe your approach.
OE_2	What difficulties or challenges did you face when using AI for this task? How did you overcome them?
OE_3	What did you learn about using AI effectively through this activity? What would you do differently next time?

References

Ansari, A. N., Ahmad, S., & Bhutta, S. M. (2024). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies, 29*(9), 11281–11321. https://doi.org/10.1007/s10639-023-12223-4

Atchley, P., Pannell, H., Wofford, K., Hopkins, M., & Atchley, R. A. (2024). Human and AI collaboration in the higher education environment: Opportunities and concerns. *Cognitive Research: Principles and Implications, 9*, Article 20. https://doi.org/10.1186/s41235-024-00547-9

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Barakat, M., Salim, N. A., & Sallam, M. (2025). University educators perspectives on ChatGPT: A technology acceptance model-based study. *Open Praxis, 17*(1), 129–144. https://doi.org/10.55982/openpraxis.17.1.718

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bettayeb, A. M., Abu Talib, M., Sobhe Altayasinah, A. Z., & Dakalbab, F. (2024). Exploring the impact of ChatGPT: Conversational AI in education. *Frontiers in Education, 9*, Article 1379796. https://doi.org/10.3389/feduc.2024.1379796

Boubker, O. (2024). From chatting to self-educating: Can AI tools boost student learning outcomes? *Expert Systems with Applications, 238*, Article 121820. https://doi.org/10.1016/j.eswa.2023.121820

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Budhathoki, T., Zirar, A., Njoya, E. T., & Timsina, A. (2024). ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). *Studies in Higher Education, 49*(5), 831–846. https://doi.org/10.1080/03075079.2024.2333937

Chiu, T. K. F. (2024a). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. *Educational Technology Research and Development, 72*(4), 2401–2416. https://doi.org/10.1007/s11423-024-10366-w

Chiu, T. K. F. (2024b). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence, 6*, Article 100197. https://doi.org/10.1016/j.caeai.2023.100197

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education & Teaching International, 61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Crawford, J., Cowling, M., & Allen, K. A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching and Learning Practice, 20*(3). https://doi.org/10.53761/1.20.3.02

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., & Soomro, R. B. (2024). Extended TAM based acceptance of AI-Powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon, 10*(8), Article e29317. https://doi.org/10.1016/j.heliyon.2024.e29317

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

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

Dwivedi, Y. K. (2025). Generative Artificial Intelligence (GenAI) in entrepreneurial education and practice: Emerging insights, the GAIN framework, and research agenda. *The International Entrepreneurship and Management Journal, 21*(1), 82. https://doi.org/10.1007/s11365-025-01089-2

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., … Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence, 6*, Article 100198. https://doi.org/10.1016/j.caeai.2023.100198

Eysenbach, G. (2023). The role of ChatGPT, generative Language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education, 9*, Article e46885. https://doi.org/10.2196/46885

Gonsalves, C. (2026). Generative AI's impact on critical thinking: Revisiting Bloom's taxonomy. *Journal of Marketing Education, 48*(1), 4–19. https://doi.org/10.1177/02734753241305980

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

Hmoud, M., & Shaqour, A. (2024). AIEd Bloom's taxonomy: A proposed model for enhancing educational efficiency and effectiveness in the artificial intelligence era. *The International Journal of Technologies in Learning, 31*(2), 111–128. https://doi.org/10.18848/2327-0144/CGP/v31i02/111-128

Hui, E. S. Y. E. (2025). Incorporating Bloom's taxonomy into promoting cognitive thinking mechanism in artificial intelligence-supported learning environments. *Interactive Learning Environments, 33*(2), 1087–1100. https://doi.org/10.1080/10494820.2024.2364237

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Ivanov, S., & Soliman, M. (2023). Game of algorithms: ChatGPT implications for the future of tourism education and research. *Journal of Tourism Futures, 9*(2), 214–221. https://doi.org/10.1108/JTF-02-2023-0038

Järvelä, S., Zhao, G., Nguyen, A., & Chen, H. (2025). Hybrid intelligence: Human–AI coevolution and learning. *British Journal of Educational Technology, 56*, 455–468. https://doi.org/10.1111/bjet.13560

Kamarova, S., Gagné, M., Holtrop, D., & Dunlop, P. D. (2025). Integrating behavior and organizational change literatures to uncover crucial psychological mechanisms underlying the adoption and maintenance of organizational change. *Journal of Organizational Behavior, 46*(2), 263–287. https://doi.org/10.1002/job.2832

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, H.-Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics, 38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52

Kim, H. K., Roknaldin, A., Nayak, S. P., Zhang, X., Twyman, M., Hwang, A. H.-C., & Lu, S. (2024). Empowering computer-supported collaborative learning with ChatGPT: Investigating effects on student interactions. *Proceedings of the 2024 ASEE PSW Conference*. https://doi.org/10.18260/1-2–46032

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

Landers, R. N., Bauer, K. N., & Callan, R. C. (2017). Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior, 71*, 508–515. https://doi.org/10.1016/j.chb.2015.08.008

Lebiere, C., Blaha, L. M., Fallon, C. K., & Jefferson, B. (2021). Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and Artificial Intelligence, 8*, Article 652776. https://doi.org/10.3389/frobt.2021.652776

Lee, C., & Cha, K. (2025). Toward the dynamic relationship between AI transparency and trust in AI: A case Study on ChatGPT. *International Journal of Human-Computer Interaction, 41*(13), 8086–8103. https://doi.org/10.1080/10447318.2024.2405266

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 CHI conference on human factors in computing systems*, Article 1121. https://doi.org/10.1145/3706598.3713778

Liang, J., Wang, L., Luo, J., Yan, Y., & Fan, C. (2023). The relationship between student interaction with generative artificial intelligence and learning achievement: Serial mediating roles of self-efficacy and cognitive engagement. *Frontiers in Psychology, 14*, Article 1285392. https://doi.org/10.3389/fpsyg.2023.1285392

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*, Article 9. https://doi.org/10.1186/s40561-024-00295-9

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Perifanou, M., & Economides, A. A. (2025). Students collaboratively prompting ChatGPT. *Computers, 14*(5), Article 156. https://doi.org/10.3390/computers14050156

Pinski, M., & Benlian, A. (2024). AI literacy for users—A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100062. https://doi.org/10.1016/j.chbah.2024.100062

Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research, 63*(7–8), 1782–1818. https://doi.org/10.1177/07356331251365189

Şahin, F., & Yıldız, G. (2024). Understanding mobile learning acceptance among university students with special needs: An exploration through the lens of self-determination theory. *Journal of Computer Assisted Learning, 40*(4), 1838–1851. https://doi.org/10.1111/jcal.12986

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th international conference on intelligent user interfaces*. https://doi.org/10.1145/3581641.3584066

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction, 39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Shahzad, M. F., Xu, S., & Asif, M. (2025). Factors affecting generative artificial intelligence, such as ChatGPT, use in higher education: An application of technology acceptance model. *British Educational Research Journal, 51*(2), 489–513. https://doi.org/10.1002/berj.4084

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Siu, B., & White, J. (2025). When ease becomes a barrier: What influences student intentions to use Generative AI in higher education? *British Educational Research Journal, 51*(6), 2893–2919. https://doi.org/10.1002/berj.4206

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

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

Sun, P. P., & Luo, X. (2024). Understanding english-as-a-foreign-language university teachers' synchronous online teaching satisfaction: A Chinese perspective. *Journal of Computer Assisted Learning, 40*(2), 685–696. https://doi.org/10.1111/jcal.12891

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

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

Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications, 13*(1), Article 684. https://doi.org/10.1057/s41599-026-07019-z

Wyszynski, M., Weber, S., Fritzsche, R., Hofgesang, M., & Niehaves, B. (2026). Satisficing vs. maximizing in prompt writing: Trait and task effects in human–AI interaction. In *Proceedings of the 2026 CHI conference on human factors in computing systems*, Article 1238. https://doi.org/10.1145/3772318.3790693

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, Article 437. https://doi.org/10.1145/3544548.3581388