*Applied Cognitive Psychology*

WILEY

RESEARCH ARTICLE OPEN ACCESS

# The Thinking Behind the Prompt: Evidence From Collaborative Student Problem Solving With Generative Artificial Intelligence

Athanasios Christopoulos[1] | Stylianos Mystakidis[2] | Isidoros Perikos[3] | Aikaterini Bourazeri[4] | Andria Michael[5] | Mikko-Jussi Laakso[1]

[1]Turku Research Institute for Learning Analytics, Faculty of Science, University of Turku, Turku, Finland | [2]School of Natural Sciences, University of Patras, Rio, Greece | [3]Computer Engineering and Informatics Department, University of Patras, GR-26504 Rio, Greece & Computer Technology Institute and Press "Diophantus", Patras, Greece | [4]School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK | [5]InSPIRE Research Centre, University of Central Lancashire Cyprus, 12–14 University Avenue, Pyla, Cyprus

**Correspondence:** Athanasios Christopoulos (atchri@utu.fi)

**Received:** 13 April 2026 | **Revised:** 1 October 2026 | **Accepted:** 9 October 2026

**Keywords:** cognitive reflection | generative artificial intelligence | human-AI interaction | intolerance of uncertainty | problem solving | self-efficacy

## ABSTRACT

Generative Artificial Intelligence (GenAI) tools are permeating education but whether cognitive dispositions moderate their benefits remains unclear. In the present work, we examined whether Cognitive Reflection and Intolerance of Uncertainty moderate the relationship between prior GenAI experience and collaborative task effectiveness. Specifically, secondary ($n = 114$) and tertiary ($n = 57$) students, distributed across groups, completed a mystery-solving task with unrestricted access to a GenAI tool of their choice. Person-centred analysis identified no cognitive-attitudinal profiles as the solution collapsed onto three groups defined by the Cognitive Reflection Test (CRT) score. In exploratory comparisons, the High-CRT group reported greater post-task prompting self-efficacy than the Low-CRT group, despite no detectable difference in performance. Cognitive reflection did not moderate the experience-effectiveness link at the group level and uncertainty intolerance did not predict AI reliance behaviours. Therefore, cognitive dispositions appear to relate more closely to the subjective experience of working with AI than to objective task outcomes.

## 1 | Introduction

Generative Artificial Intelligence (GenAI) tools are increasingly utilised in educational contexts with meta-analytic reviews reporting medium-to-large effects on academic achievement (Deng et al. 2025; Sun and Zhou 2024) and individual studies reporting positive effects on creative cognition and learner self-efficacy (Hwang and Wu 2025; Kittredge et al. 2025). However, aggregate estimates of this nature conceal considerable variation between individual users as not all of them benefit equally from Artificial Intelligence (AI) augmentation. Variability in AI-assisted outcomes has been attributed to differences in information processing, tolerance for ambiguity and willingness to accept automated outputs at face value (Fernandes et al. 2026). Hou et al. (2025) distinguished four reliance patterns among undergraduates (i.e., reflective, cautious, collaborative and thoughtless) and reported that critical thinking positively predicted the first three whereas trust predicted thoughtless use. The aforementioned findings point to an unresolved question concerning whether cognitive dispositions shape how users interact with AI or simply influence how they perceive it.

A non-negligible proportion of users appears to accept AI-generated outputs without subjecting them to critical scrutiny. Fan, Tang, et al. (2025) characterised this tendency as 'metacognitive laziness' and presented evidence that ChatGPT use raised

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Applied Cognitive Psychology* published by John Wiley & Sons Ltd.

essay scores while returning no detectable gains in knowledge acquisition or transfer. Performance that improves without a corresponding deepening of understanding is, by definition, superficial as it is disconnected from the learning processes that educational practice aims to cultivate.

Dual-process theory offers a lens through which the aforementioned variation can be examined. At the heart of the framework lies the distinction between rapid, automatic cognition (Type 1) and slower, effortful, processing (Type 2). Its relevance to AI interaction stems from the differences that exist between individuals in how willingly they accept AI-generated outputs instead of subjecting them to critical appraisal (Evans and Stanovich 2013; Brady et al. 2025). Within this broader framework, cognitive reflection occupies a central place. Frederick (2005) and Toplak et al. (2011) characterised cognitive reflection as a stable trait predictive of rational decision-making. The construct is neither identical to general intelligence nor reducible to numerical ability; it captures the disposition to override an initial, intuitively appealing answer upon recognising that it is wrong. An adversarial collaboration subsequently confirmed the Cognitive Reflection Test (CRT) as tapping a genuine cognitive trait whose validity extends beyond the mathematical domain (Meyer et al. 2024). On these grounds, individuals who score highly on the CRT would be expected to approach AI in a more strategically calibrated manner (i.e., being better positioned to discriminate between outputs that can be accepted and those that require scepticism). Low scorers, on the other hand, would be more susceptible to what the automation literature has termed as 'complacency' (Parasuraman and Manzey 2010; Wingerter et al. 2025). Cognitive reflection alone, however, cannot account for the full extent of the documented variation.

A second individual-difference variable relevant to the present inquiry is the Intolerance of Uncertainty (IU). The construct captures a dispositional tendency to experience ambiguous or uncertain situations as threatening which, in turn, generates anxiety and can motivate either vigorous information-seeking or avoidance altogether (Carleton et al. 2007; Buhr and Dugas 2002). The construct is not confined to clinical or high-stakes settings. Contemporary accounts treat it as a trans-diagnostic disposition that shapes responses to everyday uncertainty (Carleton 2016) with trait-level differences driving information seeking under such ordinary conditions (Rosen and Knäuper 2009). An information-search task built around ambiguous evidence therefore offers a legitimate arena for an a priori test. Given that AI-generated responses are typically delivered in a confident and seemingly definitive register, high-IU individuals may be drawn to such responses as a form of relief from uncertainty-related distress, even as they remain uneasy about whether the confidence is warranted. In support of this reasoning, Choi et al. (2025) reported that elevated trust diminished participants' willingness to verify AI outputs in academic settings; a finding consistent with the view that uncertainty tolerance shapes verification patterns.

A parallel body of evidence confirms that individual differences matter for AI interaction. Personality traits, cognitive styles and attitudes have each been linked to how people adopt and use AI tools (Kaya et al. 2024; Schepman and Rodway 2020, 2023). In the same vein, Fernandes et al. (2026) examined metacognitive calibration and Yang et al. (2025) traced self-efficacy development in AI-assisted settings. However, a methodological limitation runs through nearly all published studies in this area; they rely almost exclusively on variable-centred approaches which, by definition, presuppose independent and linear predictor effects. Such assumptions risk obscuring the trait configurations that, when they co-occur within a single individual, may operate in qualitatively distinct ways (Şahin and Yıldırım 2024). One finding from Fernandes et al. (2026) is especially instructive in this regard; higher AI literacy was associated with worse metacognitive calibration.

Despite the progress outlined above, whether cognitive reflection and uncertainty tolerance serve as boundary conditions to determine who benefits from AI augmentation and who does not has received insufficient attention. Person-centred approaches, such as Latent Profile Analysis (LPA), can detect naturally occurring cognitive-attitudinal configurations that variable-centred methods may overlook, but they remain largely unexplored in the AI interaction literature (Spurk et al. 2020; Ishmuradova et al. 2025; Zhang et al. 2025). The value of such methods was demonstrated by Otto et al. (2024) in the context of medical students' learning strategies where person-centred findings both extended and, on occasion, contradicted what variable-centred analyses had indicated.

### 1.1 | The Present Study

Against this background, the present study investigates how cognitive dispositions (i.e., cognitive reflection and intolerance of uncertainty) relate to collaborative problem-solving with GenAI. The inquiry is twofold: (1) a person-centred component employs LPA to examine whether cognitive-attitudinal configurations correspond to qualitatively distinct patterns of AI-augmented performance and post-task self-efficacy and (2) a variable-centred component tests whether the two dispositions moderate the relationship between prior AI experience and task effectiveness. Specifically, we investigate (a) whether individuals high in cognitive reflection engage with AI in a more strategic fashion, (b) whether uncertainty tolerance shapes trust and verification behaviours, (c) whether CRT moderates the experience-effectiveness link, (d) whether overconfidence in one's AI capability carries performance costs and (e) whether cognitive profiles differentially predict AI-assisted outcomes. By bringing dual-process theory into contact with person-centred analytical methods, the study seeks to extend the current understanding of the cognitive prerequisites for effective Human-AI collaboration. Because the task is collaborative, prompting behaviour and task effectiveness arise only at the group level.

In order to contextualise the study and structure the investigation, the following research hypotheses were formed and examined:

**H1.** *Higher cognitive reflection will be associated with more strategic AI use, operationalised as a higher proportion of higher-order prompts (*i.e., *evaluative and metacognitive) relative to total prompting behaviour.*

Individuals who score highly on the CRT are, by definition, more disposed towards Type 2 analytical processing (Evans and Stanovich 2013; Frederick 2005). The ability to judge when a simple query falls short and when a more precisely crafted prompt is required reflects the metacognitive discrimination that such processing affords. In support, Yang et al. (2025) identified 'strategic inquirers' whose structured query sequences suggest that prompting sophistication reflects, at least partially, a stable cognitive disposition.

**H2.** *Higher intolerance of uncertainty will be associated with (a) higher trust in AI outputs and (b) lower verification behaviour.*

Individuals who score highly on the IU scale experience anxiety, are more motivated to reduce that anxiety when confronted with ambiguous situations (Carleton et al. 2007). AI-generated responses, which tend to project confidence regardless of accuracy, may offer such individuals a form of psychological relief that predisposes them to place unwarranted trust in the system's outputs. At the same time, verification would reintroduce the discomfort the individual was seeking to escape.

**H3.** *Cognitive reflection will moderate the relationship between AI experience and task effectiveness such that the positive effect of experience is stronger for individuals high in CRT.*

High-CRT individuals may derive greater benefit from their prior encounters with AI because the analytical orientation that defines them (i.e., the tendency to pause, reconsider and revise) is well suited to retrospective learning from experience (Toplak et al. 2011). In the absence of such metacognitive awareness, suboptimal interaction habits may persist uncorrected.

**H4.** *Self-assessed AI capability will demonstrate a curvilinear relationship with task effectiveness such that moderate confidence optimises performance whereas overconfidence incurs penalties.*

The metacognitive calibration literature suggests that moderate self-assessed capability reflects accurate calibration which motivates appropriate verification whereas excessive confidence may discourage critical evaluation of system outputs (Fernandes et al. 2026). In line with this, metacognitive bias predicted the inclusion of incorrect AI-generated content in academic writing (Urban et al. 2025).

**H5.** *Cognitive reflection will moderate the relationship between task effectiveness and post-task self-efficacy with high-CRT individuals exhibiting stronger self-efficacy calibration following successful performance.*

According to Social Cognitive theory, mastery experiences constitute the most potent source of efficacy beliefs (Bandura 1997; Täschner et al. 2025). In contexts where AI augments human performance, the question of attribution becomes ambiguous. High-CRT individuals may be better positioned to disentangle their own contributions from those of the AI and to update their efficacy beliefs accordingly. Abdelhalim (2024) reported that high metacognitive awareness led students to use ChatGPT for stimulating thought whereas low awareness led to replication of its outputs.

**H6.** *Latent Profile Analysis will identify distinct cognitive-attitudinal profiles based on CRT, IU and GenAI attitudes, with the profiles differentially predicting strategic prompting behaviour and task effectiveness.*

Instead of presupposing that cognitive reflection, uncertainty tolerance and AI attitudes exert their effects independently and linearly, LPA can identify naturally occurring subgroups whose configurations of AI engagement constitute distinct typologies; typologies that linear models, by their very construction, cannot recover (Spurk et al. 2020; Ishmuradova et al. 2025).

## 2 | Materials and Methods

### 2.1 | Research Design

A cross-sectional observational study was conducted to investigate how cognitive dispositions influence collaborative problem-solving with GenAI tools (Figure 1). The study employed a single-session design, lasting approximately 60–90 min, with multiple sessions conducted to accommodate the full sample.

Each session began with a brief informational presentation about GenAI and Large Language Models, delivered as an educational benefit to participants. The presentation lasted approximately 10 min and covered the GenAI tools currently available, their evolution, their importance, good practices in their use and anticipated future directions; the same slides and script were used in every session, so its content was identical across sessions. Participants then completed the pre-event psychometrics individually (Appendices A–E). Following pre-task measures, groups were formed and introduced to the problem-solving task. Facilitators undertook the role of Game Masters (GMs), acting as passive observers during the process. Groups worked collaboratively to analyse evidence, formulate prompts, evaluate AI-generated responses and synthesise their conclusions. Upon completion, groups submitted their solution through the game platform which automatically calculated effectiveness scores. Participants then completed the post-event psychometrics individually (Appendices F and G). Each group interacted with its chosen tool through a single shared chat session on one member's device. Prompts were composed jointly and entered by whichever member held the device. At the close of each session, the GMs exported each group's complete log. Accordingly, every prompt was segmented and coded using a custom framework.

### 2.2 | Participants

The sample comprised 171 participants organised into 45 groups drawn through convenience sampling (Cohen et al. 2018). Participants were recruited from one Higher Education Institution and several private supplementary education centres that provide out-of-school academic support. The sample included 103 females (60.2%) and 68 males (39.8%) with ages

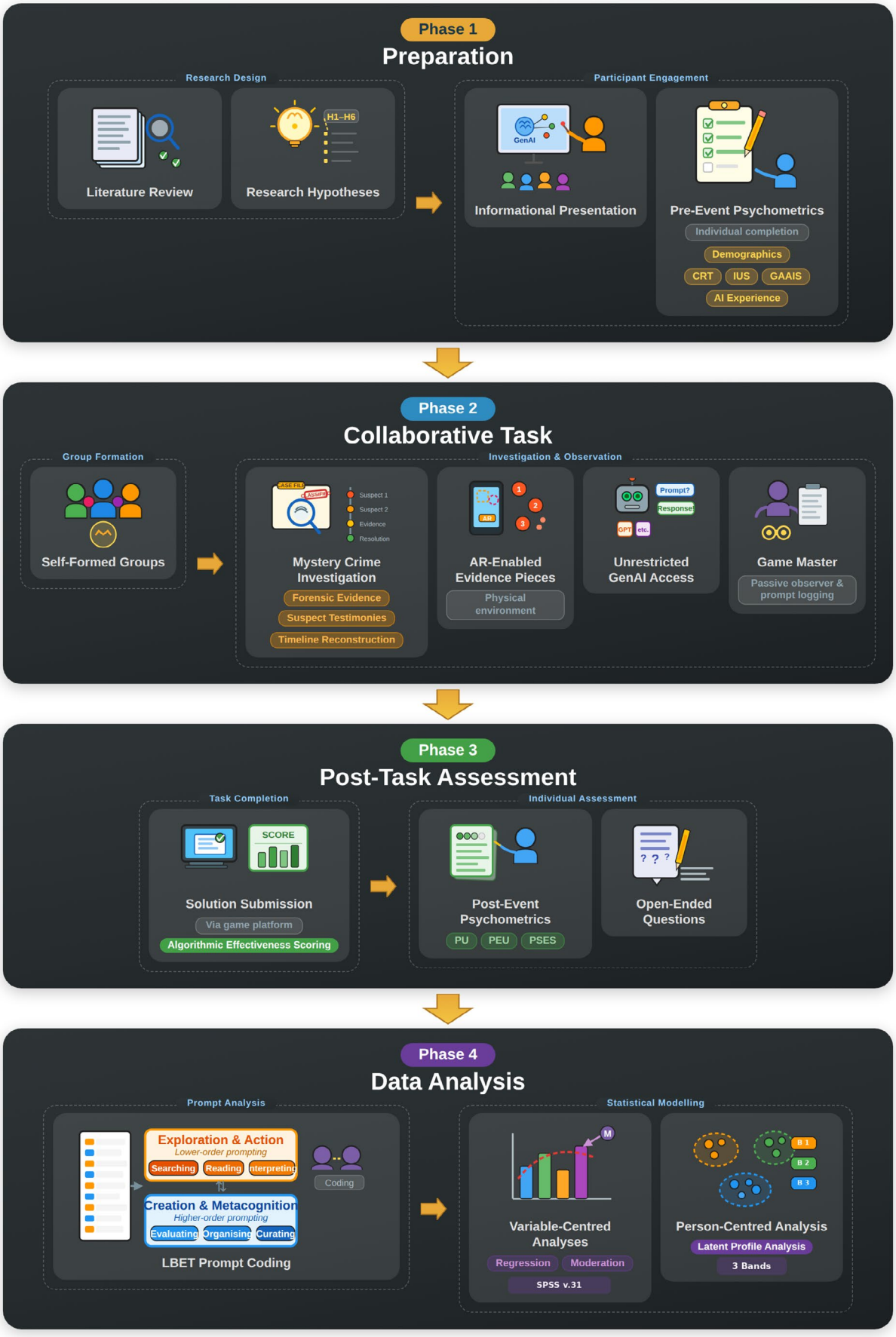



**FIGURE 1** | Overview of the research design.

ranging from 15 to 32 years (M = 18.12, SD = 3.42). Regarding the educational level, 114 participants (66.7%) were enrolled in secondary education, 49 (28.6%) 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 Informatics and Computer Engineering-related 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. The distribution of group sizes was as follows: 11 groups of three members, 32 groups of four members and two groups of five members. Group sizes in this range are standard in research on collaborative intellective problem solving (Laughlin et al. 2006; Lou et al. 1996). All participants provided informed consent prior to participation and were free to disengage at any point. Parental or guardian consent was secured for all participants below the age of 18. Participation was unremunerated, although the three top-performing teams were offered cinema tickets as a symbolic reward. The study was conducted in accordance with the Declaration of Helsinki and all data were de-identified prior to analysis.

With regard to prior exposure to GenAI, 160 out of the 171 participants (93.6%) indicated that they had used GenAI tools before taking part in the study. Duration of use varied as follows: less than 1 month ($n = 21$, 13.1%), 1–3 months ($n = 26$, 16.2%), 4–6 months ($n = 44$, 27.5%), 7–12 months ($n = 39$, 24.4%) and more than 1 year ($n = 30$, 18.8%). ChatGPT was the tool most frequently used either alone ($n = 88$) or together with Google Gemini and other tools ($n = 63$). With respect to frequency of use, the largest group reported weekly engagement (33.8%), followed by those who used AI tools two to three times per week (26.9%), whilst 16.3% reported daily use.

## 2.3 | Materials and Task

For the needs of the study, a collaborative digital mystery-solving game was developed (Figures 2 and 3). The scenario presented participants with a complex criminal investigation inquiry requiring analysis of forensic evidence, suspect testimonies and timeline reconstruction. In greater detail, groups investigated the death of a publishing executive found in his office building, with initial indications suggesting suicide by poisoning but with anomalies pointing towards potential homicide.

The case file included forensic reports, a detailed event chronology and profiles of five suspects, each with plausible motives and opportunities. Groups were required to examine 11 Augmented Reality-enabled evidence pieces distributed across the physical environment (Figure 4), evaluate 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 task was designed to require higher-order cognitive engagement in order to benefit from GenAI assistance while remaining solvable only through human reasoning and evidence integration. The difficulty of the case was calibrated through pilot testing before the main study. The competitive, leaderboard-based framing was introduced as a motivational design element, in line with evidence on the goal-setting function of leaderboards and the motivational effects of gamification in learning (Landers et al. 2017; Sailer and Homner 2020). Participants were granted the freedom to use any GenAI tool of their choice during the activity in order to capture authentic variation in tool selection and usage patterns. A total of 470 prompts were recorded across the 45 groups, with individual groups submitting between 8 and 13 prompts (M = 10.44, SD = 1.63). Task completion time ranged from 15 to 61 min (M = 32.38, SD = 12.19). Although tool choice was unrestricted, the tool each group settled on was recorded: 18 groups worked with ChatGPT, 16 with Google Gemini and 11 with Perplexity.

Mystery-solving effectiveness was computed algorithmically by the game platform based on four weighted components. The formula is as follows:

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

where $S$ represents solution accuracy (correct suspect identification, 0–1), $T$ represents completion time in minutes (capped at 90), $E$ represents the number of evidence pieces correctly interpreted (0–11) and $R$ represents the reasoning quality as rated (0–1)[1] by the GMs drawing on the problem-solving conduct recorded live during the session through structured observation. The observation rubric and the scoring anchors are reproduced in Appendix H. Effectiveness scores ranged from 0 to 100 with observed values in the sample ranging from 55 to 96. The weighting follows the logic of the task itself: identifying the correct solution is the criterial outcome and therefore carries the largest share, whereas time, evidence use and reasoning quality contribute equally to the remainder. Since any fixed weighting scheme remains, to a degree, a matter of convention, the sensitivity of the focal associations to alternative compositions of the score is examined in the Results section.

## 2.4 | Instruments

A short demographics questionnaire (Appendix A) recorded gender, age, year of study, field of study and educational level. The GenAI Experience Questionnaire (Appendix B) followed, recording the duration and frequency of GenAI use; prior AI experience was captured by a composite index derived as the sum of the duration-of-use code (0 = never used; 1 = less than 1 month; 2 = 1–3 months; 3 = 4–6 months; 4 = 7–12 months; 5 = more than 1 year) and the frequency-of-use code (0 = never; 1 = 2–3 times per month; 2 = once a week; 3 = 2–3 times per week; 4 = 4–6 times per week; 5 = daily), returning a possible range of 0–10. The same questionnaire included three single-item measures, administered before the task, assessing AI-related perceptions among experienced users: (1) self-assessed AI capability ('How would you rate your prompting capability?'), (2) trust in AI outputs ('How much do you trust GenAI

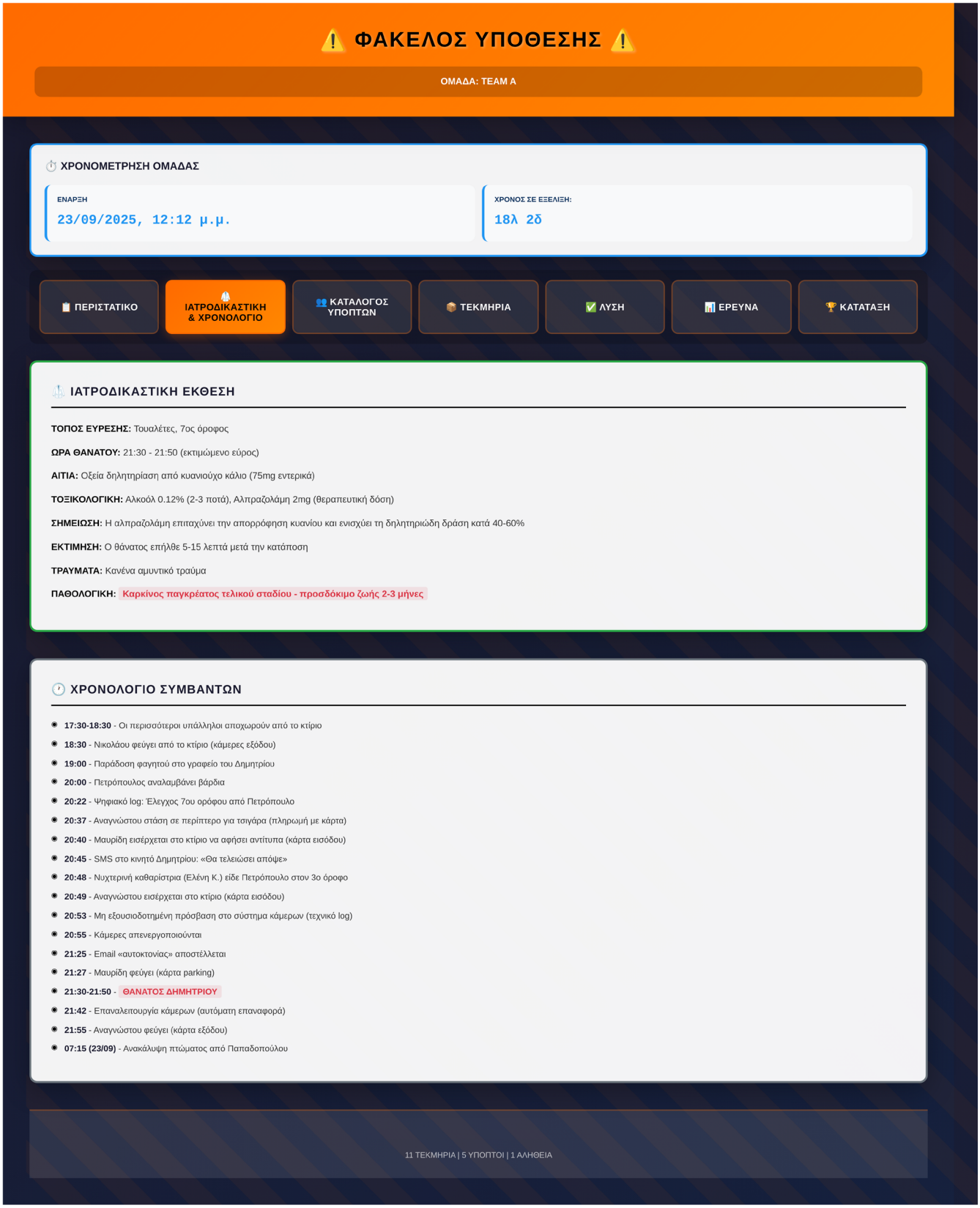



**FIGURE 2** | Player-facing case file of the Mystery Crime Investigation environment.

responses?') and (3) verification behaviour ('How often do you verify GenAI responses?'). All three items were rated on 5-point scales.

Cognitive reflection was assessed using the three-item CRT (Appendix C; Frederick 2005). Each item presents a problem for which the intuitively appealing answer is incorrect and, as

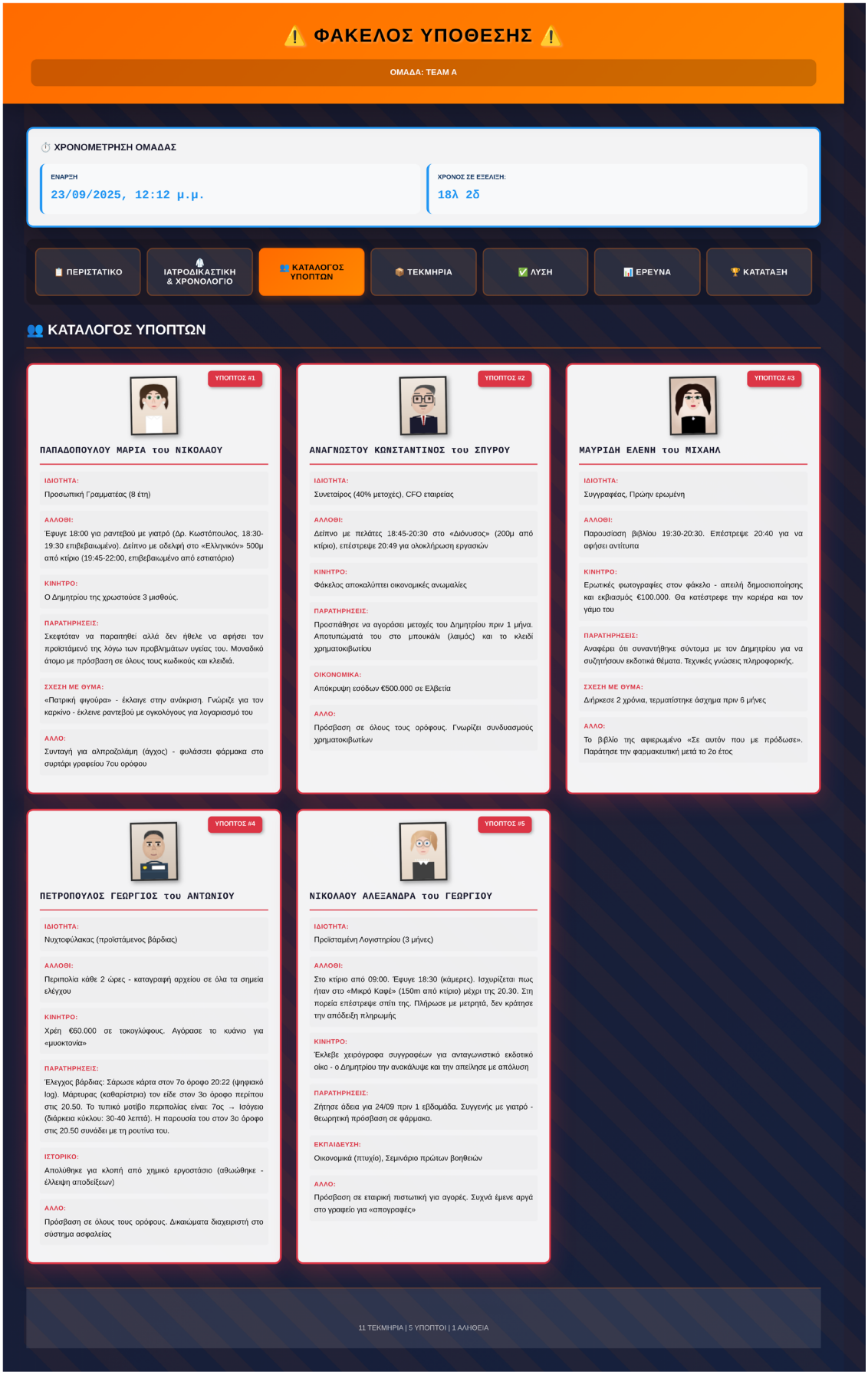



**FIGURE 3** | Suspect catalogue displaying, for each of the five suspects, their role, alibi, stated motive and investigator observations.




**FIGURE 4** | Augmented Reality evidence viewer.

such, arriving at the correct solution requires the participant to override the initial impulse and engage in analytical thinking. Scoring followed a dichotomous scheme (0 = intuitive/incorrect; 1 = reflective/correct) producing total scores from 0 to 3. Given the dichotomous scoring, internal consistency was estimated with the Kuder–Richardson formula (KR-20 = 0.40). Although this value falls below conventional expectations for Likert-based instruments, it is consistent with cognitive tests whose items are designed to tap qualitatively different aspects of a single construct instead of serving as indicators of a unitary latent factor. Comparable alpha values have been reported in validation research (Thomson and Oppenheimer 2016) and the adversarial collaboration conducted by Meyer et al. (2024) confirmed that the CRT measures a genuine thinking disposition despite modest internal consistency.

Intolerance of Uncertainty was measured with the short form six-item scale (Appendix D; Carleton et al. 2007). Items assess dispositional tendencies to experience uncertainty as aversive (e.g., 'Unforeseen events upset me greatly') with responses given on a 5-point Likert scale. The scale demonstrated good reliability ($\alpha = 0.88$).

Attitudes towards AI were measured using the eight-item General Attitudes towards Artificial Intelligence Scale (GAAIS; Appendix E) (Schepman and Rodway 2020, 2023). Five items formed the Positive subscale (e.g., 'Artificial Intelligence is exciting'; $\alpha = 0.68$) and three items the Negative subscale (e.g., 'I think Artificial Intelligence is dangerous'; $\alpha = 0.55$). All items were rated on a 5-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Both subscales therefore fall in the limited-reliability range, most likely a consequence of the abbreviated item sets, but meta-analytic evidence supports the psychometric adequacy of the GAAIS across cultural contexts (Şahin and Yıldırım 2024).

Post-task technology acceptance was assessed using the adapted Technology Acceptance Model (TAM; Appendix F) scales (Davis 1989; Lewis 2019). Perceived Usefulness (PU) was measured with six items assessing beliefs about GenAI's utility for the mystery-solving task (e.g., 'Using GenAI in this activity improved my problem-solving performance') whereas Perceived Ease of Use (PEU) was measured with six parallel items assessing effort expectations (e.g., 'I found it easy to get GenAI to do what I wanted'). All items employed a 7-point Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Internal consistency was acceptable for PU ($\alpha = 0.79$) and limited for PEU ($\alpha = 0.69$).

A custom five-item scale was developed for the present study based on Bandura's (2006) guidelines for constructing self-efficacy scales. Items of the Prompting Self-Efficacy Scale (PSES; Appendix G) 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 ($\alpha = 0.76$). An Exploratory Factor Analysis supported the unidimensionality of the scale (KMO = 0.81; Bartlett's $\chi^2(10) = 177.47$, $p < 0.001$), with a single factor accounting for 51.0% of the variance (parallel analysis likewise retained one factor), loadings between 0.63 and 0.74 and corrected item-total correlations between 0.44 and 0.56. As the scale carries no validation history, it is treated as an exploratory, study-specific measure.

Prompts were categorised using the 'Learning Behaviour with Emerging Technologies' (LBET) framework (Figure 5; Appendix I), an adaptation of Bloom's Digital Taxonomy which provides a structured approach to categorise prompts along a cognitive complexity continuum. The decision to ground the framework in Bloom's taxonomy was informed by evidence that prompt complexity tends to reflect the cognitive complexity of the questions that give rise to the prompts (Sasson Lazovsky et al. 2024). Six categories were employed: *Searching* (locating definitions and facts), *Reading* (seeking explanations or summaries), *Interpreting* (requesting examples, implications or connections), *Evaluating* (comparing, judging or critiquing information), *Organising* (synthesising and structuring material) and *Curating* (metacognitive oversight of the interaction). The categories were further aggregated into two superordinate




**FIGURE 5** | Overview of the learning behaviour with emerging technologies framework.

stages: 'Exploration and Action' (E&A; comprising *Searching, Reading* and *Interpreting*) represents lower-order prompting, characterised by information seeking and knowledge gathering, whereas 'Creation and Metacognition' (C&M; comprising *Evaluating, Organising* and *Curating*) represents higher-order prompting, characterised by synthesis, evaluation and reflective thinking. The placement of *Interpreting* within the lower-order stage is a deliberately conservative choice as requests for examples or applications do not, in themselves, require the evaluative or synthetic operations that define the higher-order stage (Hmoud and Shaqour 2024; Hui 2025).

Prompt coding was conducted by the lead researcher and verified through consensus coding. Inter-rater agreement was calculated on a stratified subset of 94 prompts (20%), independently coded by researchers of the team and resulted in a Cohen's $\kappa$ of 0.87 (excellent agreement). For the purposes of the analyses that follow, the ratio of C&M prompts to total prompts (hereafter, the C&M ratio) was treated as the principal measure of strategic AI use at the group level. The ratio indexes the prevalence of higher-order prompting within the LBET scheme; other facets of strategic engagement with AI (i.e., follow-up questioning, prompt revision, identification of AI errors and selective integration of AI-generated content) fall outside its scope and are represented in this study only by the self-reported verification measure.

## 2.5 | Data Analysis

The distributional properties of all variables were examined prior to analysis, with skewness and kurtosis evaluated against the bounds recommended by Iacobucci et al. (2025) ($|\text{skewness}| < 2$, $|\text{kurtosis}| < 7$). Because Mystery Effectiveness is computed once per group and is constant within groups, every analysis in which it appears was conducted at the group level ($N = 45$) on group-mean scores, with the corresponding degrees of freedom, in line with guidance on non-independent grouped data (Bliese 2000; Kenny et al. 2006); analyses involving only individual-level variables were conducted at the individual level. As a check on the empirical justification for aggregation, ICC(1) and ICC(2) were computed for each aggregated measure and are reported alongside the corresponding analyses.

For H1, the association between group-mean CRT and the C&M ratio was examined via group-level regression ($N = 45$). For H2, individual-level regressions explored the relationships between IU and trust/verification among the participants with prior AI experience ($n = 160$). For H3 and H5, moderation was tested at the group level through hierarchical regression on group means in which the predictor, the moderator (group-mean CRT) and their product term were entered sequentially, with all continuous variables mean-centred prior to the computation of the interaction term. For H4, a polynomial regression was conducted on group means in which the fit of a model including the quadratic term was compared against that of a linear-only specification. For H6, Latent Profile Analysis solutions specifying 2 through 5 profiles were estimated on the three dispositional indicators (CRT, IU and GAAIS-Positive). GAAIS-Negative was not among the indicators as its internal consistency ($\alpha = 0.55$) is too low to support use as a profile indicator through maximum-likelihood mixture modelling with class-invariant variances. The admissibility of each solution was inspected first and the retained classification was then

examined for differences between the CRT groups in external correlates. Prompting self-efficacy was compared across the CRT groups at the individual level ($N = 171$) using one-way ANOVA with Tukey post hoc tests. Because the C&M ratio and Mystery Effectiveness exist only at the group level—and groups contained members from more than one CRT group—each group was characterised by the proportion of its members in each CRT group and these proportions were correlated with the two group-level outcomes.

Analyses were conducted using SPSS (v.31) with the exception of the mixture estimation for H6. The alpha level was set at 0.05 unless otherwise stated. The investigation is exploratory in character and no formal correction for multiple comparisons was applied. With nine focal significance tests, a Bonferroni-adjusted criterion would set $\alpha$ at 0.0056; effects that do not meet this stricter threshold are identified as such in the Results.

**TABLE 1** | Descriptive statistics for study variables.

| Variable | M | SD | Range | $\alpha$ |
|---|---|---|---|---|
| Individual-level variables | | | | |
| CRT | 0.99 | 0.95 | 0–3 | 0.40 |
| IUS-6 | 18.47 | 4.56 | 9–29 | 0.88 |
| GAAIS-Positive | 3.38 | 0.41 | 2.2–4.6 | 0.68 |
| GAAIS-Negative | 2.89 | 0.38 | 2–4 | 0.55 |
| Perceived Usefulness | 4.80 | 0.56 | 3.2–5.8 | 0.79 |
| Perceived Ease of Use | 5.04 | 0.40 | 4.2–6.2 | 0.69 |
| Prompting Self-Efficacy | 3.95 | 0.67 | 2.4–5.6 | 0.76 |
| Experience Score | 5.63 | 2.79 | 0–10 | — |
| Group-level variables | | | | |
| Mystery Effectiveness | 78.51 | 9.14 | 55–96 | — |

## 3 | Results

### 3.1 | Descriptive Statistics

Table 1 presents the descriptive statistics and internal consistency coefficients for the study variables. The distributions were approximately normal in all cases; skewness ranged from −0.63 to 0.60 and kurtosis from −0.68 to 0.94, within the acceptable bounds. The CRT scores (range 0–3) revealed that intuitive processing predominated as 71.9% of participants scored 0–1 while only 28.1% scored 2–3. The pattern is consistent with other studies (Frederick 2005; Meyer et al. 2024). IU scores (M = 18.47, SD = 4.56) were also comparable to normative data (Carleton et al. 2007), thus suggesting typical levels of uncertainty intolerance in the sample. Several attitude measures showed restricted dispersion (GAAIS-Positive SD = 0.41 and GAAIS-Negative SD = 0.38 on 5-point scales; PEU SD = 0.40 on a 7-point scale); range restriction of this kind attenuates the correlations these variables can display. A preliminary one-way ANOVA indicated that Mystery Effectiveness did not differ across the three tools ($F(2, 42) = 1.81$, $p = 0.176$). As such, tool choice was not considered further.

Table 2 presents the bivariate correlations among the study variables. PSES was strongly and positively correlated with Mystery Effectiveness at the group level ($r = 0.87$, $p < 0.001$) and a comparably strong association was found between PSES and PU ($r = 0.61$, $p < 0.001$). The correlations point to a close link between participants' confidence in their own prompting ability, their objective performance and their appraisal of the technology's usefulness. CRT showed a small but significant positive correlation with PSES ($r = 0.22$, $p < 0.01$) which can be regarded as an indicator that the more reflective individuals tended to report somewhat higher prompting self-efficacy though the magnitude of the association was modest.

### 3.2 | Hypothesis Testing

#### 3.2.1 | Cognitive Reflection and Strategic Prompting

H1 predicted that higher group-level cognitive reflection would be associated with a greater proportion of strategic (C&M) prompts. Group-level regression revealed no significant

**TABLE 2** | Correlations among study variables.

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. CRT | — | | | | | | | |
| 2. IUS | −0.02 | — | | | | | | |
| 3. GAAIS-P | −0.08 | 0.06 | — | | | | | |
| 4. GAAIS-N | −0.03 | −0.04 | −0.50*** | — | | | | |
| 5. PU | 0.02 | 0.08 | 0.05 | 0.01 | — | | | |
| 6. PEU | −0.02 | 0.12 | 0.07 | −0.05 | 0.14 | — | | |
| 7. PSES | 0.22** | 0.07 | 0.14 | −0.06 | 0.61*** | 0.19* | — | |
| 8. ME | 0.19 | −0.11 | −0.04 | −0.14 | 0.83*** | 0.44** | 0.87*** | — |

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

association between mean CRT and the C&M ratio ($B = 0.03$, SE = 0.05, $\beta = 0.09$, $t = 0.58$, $p = 0.566$). The model accounted for negligible variance ($R^2 = 0.008$). A sensitivity power analysis indicated that, with 45 groups and a single predictor, the design afforded 0.72 power to detect a medium effect ($f^2 = 0.15$) at $\alpha = 0.05$, reaching 80% power only for effects of $f^2 \geq 0.18$; the non-significant finding therefore speaks against effects of at least that magnitude, whereas smaller effects cannot be ruled out. Aggregation itself warrants caution. CRT showed essentially no between-group variance (ICC(1) = −0.05, ICC(2) = −0.23, $F(44, 126) = 0.81$, $p = 0.783$), so group means of CRT carry little reliable group-level information. Examination of individual prompt categories revealed no significant CRT differences in the proportions of *Evaluating*, *Organising* or *Curating* prompts. It is, therefore, concluded that the first hypothesis should be rejected: cognitive reflection, when aggregated at the group level, did not predict the sophistication of prompting strategies in collaborative AI-assisted problem solving.

#### 3.2.2 | Uncertainty Intolerance, Trust and Verification

H2 predicted that higher IU would be associated with greater trust in AI outputs and lower verification behaviour. Among the experienced users ($n = 160$), IU showed no significant association with trust in AI outputs ($B = -0.004$, SE = 0.010, $\beta = -0.03$, $t = -0.41$, $p = 0.680$) and was similarly unrelated to verification behaviour ($B = -0.02$, SE = 0.02, $\beta = -0.08$, $t = -0.99$, $p = 0.322$). Effect sizes were negligible ($r < 0.10$). Both criteria were, however, measured with single items of unknown reliability and, as such, the attenuation this entails weakens both tests. On these grounds, the second hypothesis should be rejected: dispositional uncertainty intolerance showed no detectable association with AI reliance behaviours.

#### 3.2.3 | Experience-Effectiveness Moderation

H3 predicted that CRT would moderate the experience-effectiveness relationship such that the positive effect of experience would be stronger for high-CRT groups. Because Mystery Effectiveness exists only at the group level, the moderation model was estimated on group means ($N = 45$). The model as a whole was not significant ($F(3, 41) = 0.87$, $p = 0.462$, $R^2 = 0.060$). The interaction term did not reach significance ($B = 1.70$, SE = 2.63, $t = 0.65$, $p = 0.521$, 95% CI [−3.61, 7.02]). Likewise, neither the main effect of experience ($B = 0.86$, SE = 1.02, $t = 0.84$, $p = 0.407$) nor that of group-mean CRT ($B = 3.98$, SE = 3.15, $t = 1.26$, $p = 0.213$) was significant. As such, the third hypothesis should be rejected: the benefit that AI experience confers for task effectiveness did not depend on the group's level of cognitive reflection.

#### 3.2.4 | Self-Assessed Capability and Effectiveness

H4 predicted a curvilinear relationship between self-assessed AI capability and task effectiveness. Self-assessed capability, collected before the task from members with prior GenAI experience, was aggregated to the group mean and the polynomial model was estimated at that level. The group-level correlation between mean capability and effectiveness was essentially zero ($r = -0.00$, $p = 0.997$) and adding the quadratic term did not improve the model (linear $B = -0.11$, SE = 2.91, $t = -0.04$, $p = 0.970$; quadratic $B = 5.16$, SE = 5.55, $t = 0.93$, $p = 0.358$; $\Delta R^2 = 0.020$, $\Delta F = 0.86$, $p = 0.358$). The fourth hypothesis should therefore be rejected: the data provide no evidence of a curvilinear relationship between self-assessed capability and task effectiveness as no optimal capability level can be estimated from non-significant coefficients. The capability predictor, being a single item, shares the attenuation constraint noted for H2.

#### 3.2.5 | Effectiveness-Self-Efficacy Moderation

H5 predicted that CRT would moderate the relationship between task effectiveness and post-task self-efficacy. As with H3, the model was estimated on group means. Aggregation of PSES received qualified support from its intraclass correlations (ICC(1) = 0.26, ICC(2) = 0.57, $F(44, 126) = 2.33$, $p < 0.001$) with the latter value warranting caution. The model as a whole was significant ($F(3, 41) = 47.22$, $p < 0.001$, $R^2 = 0.78$) with Mystery Effectiveness emerging as a strong concurrent correlate of group-mean PSES ($B = 0.041$, SE = 0.004, $t = 11.12$, $p < 0.001$). The main effect of group-mean CRT was not significant ($B = 0.12$, SE = 0.07, $t = 1.60$, $p = 0.118$) and neither was the interaction term ($B = 0.001$, SE = 0.010, $t = 0.13$, $p = 0.898$); the effectiveness-self-efficacy relationship was virtually identical—one standard deviation above ($B = 0.042$) and below ($B = 0.041$) the mean of group CRT. Because PSES was measured immediately after the task, the effectiveness-self-efficacy association is concurrent; the data cannot establish that performance produced the efficacy beliefs and part of the association's magnitude may reflect shared task content.

Given that the effectiveness score rests on a partly conventional weighting, the H5 model was re-estimated under alternative compositions of the score. An equal-weights recomposition and a recomposition restricted to the objective components (i.e., omitting the GM-rated reasoning term) correlated $r = 0.98$ and $r = 0.96$ with the original score, respectively. Group-mean PSES remained strongly associated with both ($r = 0.87$ and $r = 0.82$, $p < 0.001$), the effectiveness term remained significant in the moderation model under both compositions ($t = 10.63$ and $t = 9.24$, $p < 0.001$) and the interaction remained non-significant ($p = 0.834$ and $p = 0.730$). The focal association is therefore not an artefact of the weighting scheme or of the single observer-rated component. It is concluded that the fifth hypothesis should be rejected: cognitive reflection did not moderate the relationship between task effectiveness and post-task self-efficacy.

#### 3.2.6 | Person-Centred Structure

H6 predicted that LPA would identify distinct cognitive-attitudinal profiles with differential associations to prompting behaviour and task effectiveness. Mixture solutions specifying two through five profiles were estimated on the three dispositional indicators. The structure that emerged, however, was carried by cognitive reflection alone: the maximum-likelihood three-profile solution collapsed onto the observed CRT score bands, classifying participants with certainty into those

answering two or three items correctly, one item and none, with no separation on uncertainty intolerance or AI attitudes. The three profiles were therefore operationalised as CRT score groups, labelled as the 'High-CRT group' (CRT 2–3; $n = 48$, 28.1%), the 'Intermediate-CRT group' (CRT = 1; $n = 59$, 34.5%) and the 'Low-CRT group' (CRT = 0; $n = 64$, 37.4%). Across the three CRT groups, uncertainty intolerance (IU M = 18.42, 18.20 and 18.75, respectively) and positive AI attitudes (GAAIS-Positive M = 3.37, 3.33 and 3.44) were essentially flat, as was the negative-attitude subscale when inspected descriptively (GAAIS-Negative M = 2.85, 2.94 and 2.88). The mixture analysis identified no structure beyond the CRT score itself.

In exploratory comparisons, the CRT groups differed in prompting self-efficacy ($F(2, 168) = 4.65$, $p = 0.011$, $\eta^2 = 0.052$; Figure 6). The comparison was not among the pre-specified profile outcomes and does not survive the Bonferroni-adjusted criterion ($\alpha = 0.0056$). Tukey's HSD indicated that the High-CRT group (M = 4.19, SD = 0.80) scored significantly higher on the PSES than the Low-CRT group (M = 3.82, SD = 0.52, $p = 0.010$, $d = 0.57$) whereas the Intermediate-CRT group (M = 3.90, SD = 0.67) fell between the two. For the group-level outcomes, by contrast, CRT-group composition was unrelated to performance as the proportion of High-CRT participants in a collaborative group correlated with neither Mystery Effectiveness ($r = 0.18$, $p = 0.238$) nor the C&M ratio ($r = 0.13$, $p = 0.385$). The corresponding correlations for the other two CRT groups were smaller still ($|r| \leq 0.14$, $p \geq 0.351$). The CRT groups exhibited a stronger relation to the psychological dimension of AI collaboration (i.e., post-task confidence) than to the objective task outcomes. The sixth hypothesis was, therefore, not supported as no cognitive-attitudinal configuration emerged and the CRT groups did not differ in prompting behaviour or task effectiveness. The CRT-based classification is accordingly treated as exploratory.

### 3.3 | Supplementary Analyses

Because group size varied from three to five members, its potential influence was also examined. Size correlated with the number of prompts a group produced ($r = 0.39$, $p = 0.009$) and with nothing else: correlations with Mystery Effectiveness ($r = 0.11$, $p = 0.486$), the C&M ratio ($r = 0.19$, $p = 0.214$) and mean prompting self-efficacy ($r = -0.09$, $p = 0.571$) were all non-significant. Re-estimating the H3, H4 and H5 models with group size entered as a covariate left every substantive conclusion unchanged (H3: $F(4, 40) = 0.83$, $p = 0.516$, interaction $p = 0.538$; H4: quadratic $p = 0.420$; H5: $F(4, 40) = 41.98$, $R^2 = 0.81$, with the effectiveness term robust, $B = 0.042$, $t = 12.06$, $p < 0.001$ and the interaction still non-significant, $p = 0.737$). One exploratory observation emerged: adjusting for effectiveness and cognitive reflection, larger groups reported somewhat lower post-task self-efficacy ($B = -0.160$, $p = 0.014$). The coefficient does not survive the Bonferroni-adjusted criterion ($\alpha = 0.0056$) and rests on only two five-member groups.

## 4 | Discussion

The aim of the current study was to examine whether cognitive dispositions prove decisive in determining how individuals interact with AI during collaborative problem solving. On the basis of dual-process theory, we hypothesised that higher cognitive reflection would be associated with more strategic prompting and that elevated uncertainty intolerance would be associated with excessive trust in AI outputs. Neither prediction received confirmation. What emerged instead was a dissociation between process and outcome; the associations involving cognitive dispositions were confined to the subjective experience of working with AI (i.e., post-task self-efficacy) and did not extend to the observable performance indicators.

### 4.1 | Cognitive-Reflection Groups and Post-Task Self-Efficacy

The person-centred analysis identified no cognitive-attitudinal configuration; the solution reduced to the CRT score itself in a domain where person-centred questions have been underrepresented (Howard and Hoffman 2018). The finding we consider

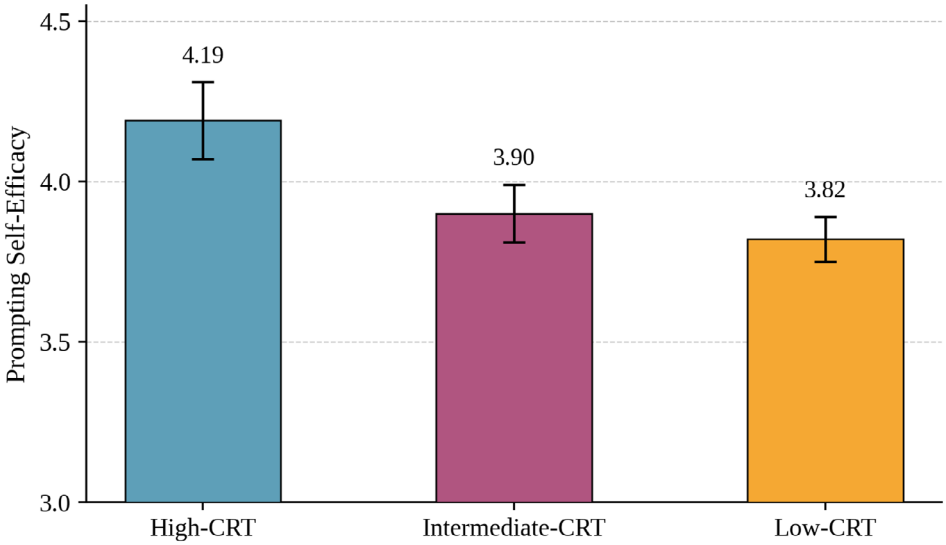



**FIGURE 6** | Mean prompting self-efficacy *by CRT group*. Error bars represent ±1 standard error; the response scale ranged from 1 to 7.

most consequential—exploratory, as the CRT-based classification was not the hypothesised profile structure and the comparison was not among the pre-specified outcomes—is the divergence between the CRT groups on self-efficacy alongside the absence of detectable performance differences. The High-CRT group expressed substantially greater confidence in their prompting abilities than the Low-CRT group despite the fact that their performance scores were not distinguishable from those of the other two CRT groups. The immediate, task-level gains that AI tools afford appear to be within reach of users across the cognitive spectrum; what differed was the subjective return participants reported after the task (i.e., the confidence they took from the collaborative experience).

Such dissociations are not without precedent. Meta-analytic work on activity achievement emotions shows reliable but modest associations between the emotional experience of learning and academic performance (Camacho-Morles et al. 2021), a pattern that treats the experience of learning as an outcome worth explaining in its own right and leaves ample room for experience and achievement to diverge. Person-centred methods have been shown to uncover differences that elude variable-centred analyses (Otto et al. 2024), although in the present study the mixture analysis separated participants along a single cognitive dimension only.

Our interpretation is that individuals with higher cognitive reflection engage with AI in a more deliberate and self-aware fashion and that this quality of engagement, instead of any demonstrable performance advantage, accounts for the observed divergence in self-efficacy. Even though their approach to the collaborative task did not produce superior outcomes, it may have given them a firmer sense of agency over the process, a more thorough assimilation of the collaborative experience and a more articulate understanding of the strategies they had deployed. The metacognitive vigilance these individuals bring to their AI interactions may also incline them to attribute successful outcomes to their own decisions and judgements as opposed to the AI's contribution; an attributional pattern that, following the logic advanced by Bandura (1997), would strengthen self-efficacy through a more precisely registered mastery experience.

The dissociation carries practical implications for AI literacy education as users whose cognitive profiles differ may need different kinds of instructional support. Learners in the Low-CRT group stand to gain from explicit scaffolding designed to foster strategic engagement and metacognitive monitoring. As Weidlich et al. (2025) argued, educational technology research ought to attend to psychological outcomes no less than to performance. The present study is in line with the aforementioned view as the CRT groups differed in psychological response in the absence of differential performance.

### 4.2 | Cognitive Reflection and Strategic AI Interaction

The absence of any association between group-level cognitive reflection and strategic prompting should be interpreted with caution. Groups with higher average CRT scores did not generate a greater proportion of evaluative, organisational or metacognitive prompts. Dual-process theory would predict otherwise, as would the findings from Meyer et al. (2024), who affirmed the CRT as a genuine thinking disposition and from Hou et al. (2025), who showed critical thinking to be predictive of reflective, cautious and collaborative AI use patterns.

As we did not anticipate this result, several considerations—albeit post hoc—may account for it. The collaborative structure of the task is perhaps most relevant. When individuals with divergent cognitive orientations are assembled within a single working group, group-level dynamics may attenuate the tendencies that any given member would manifest in isolation. A characteristically reflective participant might, for example, defer to a more assertive group member who favours a more intuitive approach. Collective-cognition research supports this interpretation as conformity pressures and shared mental models within teams suppress the behavioural signatures of individual cognitive differences (Tylén et al. 2023). Besides, Jin and Sercu (2025) reported that interventions spanning medium-term periods produced more favourable results than those confined to a single session. The brief duration of the present collaborative task may not have afforded enough time for cognitive differences to become manifest in observable prompting behaviour.

It should also be acknowledged that the LBET framework, though grounded in Bloom's revised taxonomy, may not have captured those dimensions of strategic AI use that are most sensitive to cognitive reflection. Strategy in AI interaction manifests both in the type of prompt one issues and in the iterative refinement of prompts, the critical evaluation of AI responses and the selectivity with which AI-generated content is integrated into the final product. Interview studies of GenAI-supported collaborative argumentation describe engagement of this kind. Students who argued with the support of ChatGPT reported that it assisted them in deconstructing complex problems, evaluating arguments and generating solutions, while over-reliance, cognitive overload and bias in its outputs were identified as difficulties (Fan, Kim, et al. 2025). Sstudents have likewise described the system as a tool, a facilitator, a teaching assistant and even a ‘machine buddy’ (Kim et al. 2026). Evaluating an argument the system has generated and deciding which of its suggestions the group adopts are higher-order processes that leave no trace in a count of prompt categories. The prompt log recorded what each group asked, whereas how the group treated what it received was not coded. If cognitive reflection is expressed in that treatment, the C&M ratio could not have detected it. As such, designs that pair prompt logs with argumentation records would offer a more direct test.

The absence of an association should not be taken as evidence against a theoretical link between cognitive reflection and strategic AI use. Instead, the finding suggests that the link is context-dependent—operative when tasks are open-ended and when prompting quality differentiates outcomes but dormant when the task structure scaffolds adequate performance regardless of who is doing the prompting. Measurement further constrains what these tests could show. In view of the low internal consistency, the CRT enters the models with substantial unreliability and attenuation of that magnitude mechanically reduces

the power to detect both group-level associations and interaction effects; the non-significant moderation results for H3 and H5 must be read against this constraint.

### 4.3 | Intolerance of Uncertainty and AI Reliance

The finding that IU was associated with neither trust nor verification was unexpected, given the well-established role of certainty-seeking in anxiety reduction. Two considerations, both post hoc, may account for the discrepancy. The present study was conducted in a context where incorrect answers carried no real-world penalty (c.f., participants' grades in educational settings). Under these conditions, the uncertainty-reduction motivation that typically drives IU-related behaviours may not have been activated. Instead of indicating a flaw in the task, the pattern delimits the construct's reach. Trait IU operates under everyday uncertainty (Carleton 2016; Rosen and Knäuper 2009) but its influence on reliance behaviours may become observable only when the stakes rise, as in medical decision-making, financial choices or formal examinations. In addition, the construct validity of IU in AI interaction contexts requires further examination. The 'uncertainty' generated by an AI system (i.e., the question of whether a particular response is accurate) may be qualitatively different from the interpersonal and social forms of uncertainty that the IU construct was designed to measure. If evaluating a technological system's reliability and managing social ambiguity are not the same cognitive operation, then domain-specific instruments will be necessary before AI interaction can be studied with the precision the topic demands. The findings of Choi et al. (2025), who reported that verification depended on task salience more than on dispositional factors, lend support to this interpretation.

## 5 | Implications

### 5.1 | Implications for Theory and Research

For dual-process accounts of AI interaction, the findings indicate a boundary condition. The disposition that separates reflective from intuitive thinkers left no detectable trace on collective prompting behaviour but reappeared in what individuals took from the experience. Predictions derived from individual cognition may therefore need to specify the social configuration in which they are expected to hold, with collaborative settings attenuating dispositional expression at the behavioural level (Evans and Stanovich 2013; Tylén et al. 2023).

A second implication concerns how mastery experiences are theorised when performance is achieved jointly with an AI system. Bandura (1997) tied efficacy growth to the successful execution of a task, but in the present study, comparable group performance was accompanied by different degrees of post-task confidence depending on the cognitive disposition of the member. Accounts of self-efficacy formation in AI-assisted settings (Yang et al. 2025) may accordingly need an attribution component specifying how credit for jointly produced outcomes is divided between the user and the system.

A methodological implication also emerges for educational technology research. Person-centred analyses—such as LPA—can reveal patterns that remain hidden within variable-centred frameworks. The typology of analytic thinkers that Svedholm-Häkkinen (2025) put forward (i.e., one that differentiates between Overall Analytic, Open and Reflective subtypes) may be particularly useful in determining which individuals are most likely to benefit from different forms of AI literacy instruction.

### 5.2 | Implications for Practice

That cognitive reflection did not predict strategic prompting when users worked in groups carries an informative implication of its own as it would be premature to assume that individual cognitive dispositions carry over, in any straightforward fashion, to group-level AI interaction behaviours. Educators may, therefore, find it necessary to scaffold strategic prompting explicitly for all learners regardless of cognitive profile. Structured metacognitive support and progressive prompting guidance both produce measurable gains in GenAI environments (Li et al. 2025; Xu et al. 2025). In view of such findings, AI literacy curricula should provide dedicated instruction in query formulation, critical evaluation of system responses and iterative refinement of prompts (i.e., competences that do not emerge spontaneously from cognitive ability).

The divergence between performance and self-efficacy raises a concern about how AI-assisted learning is currently evaluated. If the High-CRT group develops stronger confidence while achieving indistinguishable scores, evaluation approaches that consider only the product of learning leave out a dimension that may, over time, prove at least as consequential. Assessment of AI-assisted learning should, therefore, incorporate process-oriented measures, self-efficacy indicators and qualitative appraisals of learners' reflections on their interaction strategies in addition to conventional performance measures.

The findings for IU point to a different kind of implication. The absence of an association cannot demonstrate the absence of an effect but it is compatible with the possibility that the design of AI interfaces and of tasks constitutes a more proximal determinant of reliance behaviour than personality-level characteristics. Design features such as built-in verification prompts, explicit uncertainty markers and scaffolded evaluation procedures may accordingly offer a more direct path than interventions targeting dispositional responses to uncertainty. The three CRT groups reinforce this point as a single pedagogical strategy is unlikely to serve all learners equally. Therefore, brief cognitive assessments could inform the design of differentiated instruction, with content organised along the competency dimensions that AI-literacy frameworks describe (Long and Magerko 2020; Ng et al. 2021).

## 6 | Limitations and Future Directions

The study's limitations provide directions for future research. The first concerns the CRT. With only three items and low

internal consistency, the instrument is blunt. Meyer et al. (2024) confirmed its predictive validity, but it may lack the sensitivity needed to detect the moderation effects we hypothesised. The consequence is statistical as unreliability attenuates observed associations (Spearman 1904), with moderation tests being among the least powered designs in applied research (Aguinis et al. 2005). Given that level of unreliability, the moderation tests involving cognitive reflection were underpowered against true effects of the hypothesised form and so their non-significant results may reflect measurement attenuation (a Type II risk) as much as the absence of the effects themselves. Longer CRT variants or verbal alternatives that bypass the well-documented confound with numeracy would improve measurement precision in future investigations. Because assignment to the CRT groups is a deterministic function of the CRT score, measurement error in the CRT propagates directly into the classification itself.

Second, the cross-sectional design precludes causal claims about moderation. Longitudinal studies that track prompting behaviours and self-efficacy across repeated encounters with GenAI tools would be more informative. Molenaar et al. (2023) have also made this argument in calling for multimodal, longitudinal approaches to the study of self-regulated learning with AI.

Third, the collaborative task format introduces dependencies between individual participants that constrain to some extent the analyses. Mystery Effectiveness was computed once per group and therefore has no within-group variance by construction. Aggregating CRT and other individual-difference variables to the group level, conversely, discards within-group variation that may carry theoretically meaningful information. Group composition adds a further constraint as participants self-selected into groups. The compositional homogeneity—which may be linked to friendship ties, ability or prior GenAI experience—cannot be separated from group processes in the present design (Chapman et al. 2006). Self-selected groups do, however, tend to perform comparably to assigned groups with stronger cohesion (Hilton and Phillips 2010). Besides, naturalistic formation mirrors everyday classroom collaboration. Subsequent studies should examine AI interaction at the individual level in settings free of group dynamics or, when collaboration is part of the design, employ multilevel models that partition variance at both levels.

Fourth, several measurement constraints should be noted. The reasoning component of the effectiveness score was rated by a single GM per group. The time component may partly conflate efficient problem-solving with reduced deliberation. Trust, verification and self-assessed capability were, moreover, measured with single items. Single-item measures can perform acceptably for narrow, well-defined constructs (Allen et al. 2022; Fisher et al. 2016), but the constructs at issue here are evaluative dispositions for which multi-item instruments remain preferable. The competitive, leaderboard-based framing of the task may likewise have shaped prompting behaviour in ways this design cannot separate from the task itself. The standardised pre-task presentation, albeit identical in every session, may in addition have shifted the absolute levels of the attitude and efficacy measures.

Fifth, generalisability warrants caution. The participants were secondary- and tertiary-level students, predominantly female, with relatively similar prior AI exposure. Whether the patterns observed would replicate among working professionals whose GenAI experience is more extensive and for whom errors carry real consequences remains an open question. Most GenAI research has occurred in Higher Education (Yusuf et al. 2024). By including secondary-level participants, the present study begins to address this gap, although the data do not permit generalisation to professional or workplace contexts.

Finally, the single-session format may not have afforded sufficient time for cognitive profiles to give rise to distinguishable developmental pathways. Bandura (1997) theorised that self-efficacy builds through cumulative mastery experiences. The CRT-group differences detected after a single task could either widen or vanish with repetition—a question only longitudinal tracking can resolve. Genuine but small effects may have gone undetected as the group-level tests afford limited power against medium-sized effects. Tasks that foreground active knowledge construction may also be more conducive to surfacing cognitive differences in AI-assisted learning (Chen et al. 2026; Liu and Wang 2024).

## 7 | Conclusion

In the present work we asked whether cognitive dispositions determine how collaborative groups work with GenAI and what they take away from the experience. At the group level, the answer for objective outcomes is negative; no disposition predicted strategic prompting or reliance behaviours, none moderated the experience-effectiveness link and self-assessed capability showed no relation to task effectiveness. The person-centred analysis found no cognitive-attitudinal configuration beyond the CRT score itself. In exploratory comparisons, the three CRT groups differed in post-task prompting self-efficacy even though their performance was indistinguishable. In collaborative settings of this kind, cognitive dispositions therefore appear to bear more on how learners experience AI-assisted work than on how well their groups perform. For practice, the findings support cognitively differentiated AI literacy instruction. For research, they argue for designs that can trace individual prompting behaviour and its development over time.

### Author Contributions

**Athanasios Christopoulos:** conceptualization, investigation, methodology, formal analysis, data curation, software, writing – original draft, project administration, funding acquisition, visualization. **Stylianos Mystakidis:** conceptualization, methodology, writing – review and editing, resources. **Isidoros Perikos:** validation, writing – review and editing, resources. **Aikaterini Bourazeri:** data curation, writing – review and editing. **Andria Michael:** writing – review and editing. **Mikko-Jussi Laakso:** funding acquisition, supervision, writing – review and editing.

### Acknowledgements

The authors wish to thank all participating students and volunteers who contributed their time to support the study. Open access publishing facilitated by Turun yliopisto, as part of the Wiley - FinELib agreement.

**Funding**

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

**Ethics Statement**

The study was conducted in accordance with the ethical standards of the Declaration of Helsinki (1964) and its later amendments. All participants provided informed consent 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.

**Conflicts of Interest**

The authors declare no conflicts of interest.

**Data Availability Statement**

The dataset is available from the corresponding author upon reasonable request.

**Endnotes**

[1] No coherent reasoning = 0; Fully coherent, evidence-integrated, chain of inference = 1.

**References**

Abdelhalim, S. M. 2024. "Using ChatGPT to Promote Research Competency: English as a Foreign Language Undergraduates' Perceptions and Practices Across Varied Metacognitive Awareness Levels." *Journal of Computer Assisted Learning* 40, no. 3: 1261–1275. https://doi.org/10.1111/jcal.12948.

Aguinis, H., J. C. Beaty, R. J. Boik, and C. A. Pierce. 2005. "Effect Size and Power in Assessing Moderating Effects of Categorical Variables Using Multiple Regression: A 30-Year Review." *Journal of Applied Psychology* 90, no. 1: 94–107. https://doi.org/10.1037/0021-9010.90.1.94.

Allen, M. S., D. Iliescu, and S. Greiff. 2022. "Single Item Measures in Psychological Science: A Call to Action [Editorial]." *European Journal of Psychological Assessment* 38, no. 1: 1–5. https://doi.org/10.1027/1015-5759/a000699.

Bandura, A. 1997. *Self-Efficacy: The Exercise of Control*. W. H. Freeman.

Bandura, A. 2006. "Guide for Constructing Self-Efficacy Scales." In *Self-Efficacy Beliefs of Adolescents*, edited by F. Pajares and T. Urdan, vol. 5, 307–337. Information Age Publishing.

Bliese, P. D. 2000. "Within-Group Agreement, Non-Independence, and Reliability: Implications for Data Aggregation and Analysis." In *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, edited by K. J. Klein and S. W. J. Kozlowski, 349–381. Jossey-Bass.

Brady, O., P. Nulty, L. Zhang, T. E. Ward, and D. P. McGovern. 2025. "Dual-Process Theory and Decision-Making in Large Language Models." *Nature Reviews Psychology* 4: 777–792. https://doi.org/10.1038/s44159-025-00506-1.

Buhr, K., and M. J. Dugas. 2002. "The Intolerance of Uncertainty Scale: Psychometric Properties of the English Version." *Behaviour Research and Therapy* 40, no. 8: 931–945. https://doi.org/10.1016/S0005-7967(01)00092-4.

Camacho-Morles, J., G. R. Slemp, R. Pekrun, K. Loderer, H. Hou, and L. G. Oades. 2021. "Activity Achievement Emotions and Academic Performance: A Meta-Analysis." *Educational Psychology Review* 33: 1051–1095. https://doi.org/10.1007/s10648-020-09585-3.

Carleton, R. N. 2016. "Fear of the Unknown: One Fear to Rule Them All?" *Journal of Anxiety Disorders* 41: 5–21. https://doi.org/10.1016/j.janxdis.2016.03.011.

Carleton, R. N., P. J. Norton, and G. J. G. Asmundson. 2007. "Fearing the Unknown: A Short Version of the Intolerance of Uncertainty Scale." *Journal of Anxiety Disorders* 21, no. 1: 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014.

Chapman, K. J., M. Meuter, D. Toy, and L. Wright. 2006. "Can't We Pick Our Own Groups? The Influence of Group Selection Method on Group Dynamics and Outcomes." *Journal of Management Education* 30, no. 4: 557–569. https://doi.org/10.1177/1052562905284872.

Chen, A., Y. Wei, H. Le, and Y. Zhang. 2026. "Learning by Teaching With ChatGPT: The Effect of Teachable ChatGPT Agent on Programming Education." *British Journal of Educational Technology* 57: 163–184. https://doi.org/10.1111/bjet.70001.

Choi, W., H. Bak, J. An, Y. Zhang, and B. Stvilia. 2025. "College Students' Credibility Assessments of GenAI-Generated Information for Academic Tasks: An Interview Study." *Journal of the Association for Information Science and Technology* 76, no. 6: 867–883. https://doi.org/10.1002/asi.24978.

Cohen, L., L. Manion, and K. Morrison. 2018. *Research Methods in Education*. 8th ed. Routledge.

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." *MIS Quarterly* 13, no. 3: 319–340. https://doi.org/10.2307/249008.

Deng, R., M. Jiang, X. Yu, Y. Lu, and S. Liu. 2025. "Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-Analysis of Experimental Studies." *Computers & Education* 227: 105224. https://doi.org/10.1016/j.compedu.2024.105224.

Evans, J., and K. E. Stanovich. 2013. "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science* 8, no. 3: 223–241. https://doi.org/10.1177/1745691612460685.

Fan, L., J. Kim, R. Detrick, and N. Li. 2025. "Developing Students' Higher-Order Thinking Skills With Generative AI Insights and Strategies From a Case Study." In *Prompt Engineering and Generative AI Applications for Teaching and Learning*, edited by A. ElSayary, 411–428. IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-7332-3.ch024.

Fan, Y., L. Tang, H. Le, et al. 2025. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology* 56: 489–530. https://doi.org/10.1111/bjet.13544.

Fernandes, D., S. Villa, S. Nicholls, et al. 2026. "AI Makes You Smarter but None the Wiser: The Disconnect Between Performance and Metacognition." *Computers in Human Behavior* 175: 108779. https://doi.org/10.1016/j.chb.2025.108779.

Fisher, G. G., R. A. Matthews, and A. M. Gibbons. 2016. "Developing and Investigating the Use of Single-Item Measures in Organizational Research." *Journal of Occupational Health Psychology* 21, no. 1: 3–23. https://doi.org/10.1037/a0039139.

Frederick, S. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19, no. 4: 25–42. https://doi.org/10.1257/089533005775196732.

Hilton, S., and F. Phillips. 2010. "Instructor-Assigned and Student-Selected Groups: A View From Inside." *Issues in Accounting Education* 25, no. 1: 15–33. https://doi.org/10.2308/iace.2010.25.1.15.

Hmoud, M., and A. Shaqour. 2024. "AIEd Bloom's Taxonomy: A Proposed Model for Enhancing Educational Efficiency and Effectiveness in the Artificial Intelligence Era." *International Journal of Technologies in Learning* 31, no. 2: 111–128. https://doi.org/10.18848/2327-0144/CGP/v31i02/111-128.

Hou, C., G. Zhu, and V. Sudarshan. 2025. "The Role of Critical Thinking on Undergraduates' Reliance Behaviours on Generative AI in Problem-Solving." *British Journal of Educational Technology* 56, no. 5: 1919–1941. https://doi.org/10.1111/bjet.13613.

Howard, M. C., and M. E. Hoffman. 2018. "Variable-Centered, Person-Centered, and Person-Specific Approaches: Where Theory Meets the Method." *Organizational Research Methods* 21, no. 4: 846–876. https://doi.org/10.1177/1094428117744021.

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, no. 2: 1–14. https://doi.org/10.1080/10494820.2024.2364237.

Hwang, Y., and Y. Wu. 2025. "The Influence of Generative Artificial Intelligence on Creative Cognition of Design Students: A Chain Mediation Model of Self-Efficacy and Anxiety." *Frontiers in Psychology* 15: 1455015. https://doi.org/10.3389/fpsyg.2024.1455015.

Iacobucci, D., S. Román, S. Moon, and D. Rouziès. 2025. "A Tutorial on What to Do With Skewness, Kurtosis, and Outliers: New Insights to Help Scholars Conduct and Defend Their Research." *Psychology & Marketing* 42, no. 5: 1398–1414. https://doi.org/10.1002/mar.22187.

Ishmuradova, I. I., A. A. Chistyakov, T. A. Brodskaya, N. N. Kosarenko, N. V. Savchenko, and N. N. Shindryaeva. 2025. "Latent Profiles of AI Learning Conditions Among University Students: Implications for Educational Intentions." *Contemporary Educational Technology* 17, no. 2: ep565. https://doi.org/10.30935/cedtech/15907.

Jin, Y., and L. Sercu. 2025. "ChatGPT Interventions in Higher Education: A Systematic Review of Experimental Studies." *Journal of Computer Assisted Learning* 41, no. 4: e70072. https://doi.org/10.1111/jcal.70072.

Kaya, F., F. Aydin, A. Schepman, P. Rodway, O. Yetisensoy, and M. Demir Kaya. 2024. "The Roles of Personality Traits, AI Anxiety, and Demographic Factors in Attitudes Toward Artificial Intelligence." *International Journal of Human-Computer Interaction* 40, no. 2: 497–514. https://doi.org/10.1080/10447318.2022.2151730.

Kenny, D. A., D. A. Kashy, and W. L. Cook. 2006. *Dyadic Data Analysis*. Guilford Press.

Kim, J., S. Yu, R. Detrick, L. Fan, and N. Li. 2026. "Students' Perception of Generative AI-Assisted Collaborative Argumentation." *European Journal of Education* 61, no. 2: e70576. https://doi.org/10.1111/ejed.70576.

Kittredge, A. K., E. W. M. Hopman, B. Reuveni, D. Dionne, C. Freeman, and X. Jiang. 2025. "Mobile Language App Learners' Self-Efficacy Increases After Using Generative AI." *Frontiers in Education* 10: 1499497. https://doi.org/10.3389/feduc.2025.1499497.

Landers, R. N., K. N. Bauer, and R. C. Callan. 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.

Laughlin, P. R., E. C. Hatch, J. S. Silver, and L. Boh. 2006. "Groups Perform Better Than the Best Individuals on Letters-To-Numbers Problems: Effects of Group Size." *Journal of Personality and Social Psychology* 90, no. 4: 644–651. https://doi.org/10.1037/0022-3514.90.4.644.

Lewis, J. R. 2019. "Comparison of Four TAM Item Formats: Effect of Response Option Labels and Order." *Journal of Usability Studies* 14, no. 4: 224–236. https://doi.org/10.5555/3542805.3542809.

Li, C., G. Hwang, C. Chang, and H. Su. 2025. "Generative AI-Supported Progressive Prompting for Professional Training: Effects on Learning Achievement, Critical Thinking, and Cognitive Load." *British Journal of Educational Technology* 56, no. 6: 2550–2572. https://doi.org/10.1111/bjet.13594.

Liu, W., and Y. Wang. 2024. "The Effects of Using AI Tools on Critical Thinking in English Literature Classes Among EFL Learners: An Intervention Study." *European Journal of Education* 59, no. 4: e12804. https://doi.org/10.1111/ejed.12804.

Long, D., and B. Magerko. 2020. "What Is AI Literacy? Competencies and Design Considerations." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. ACM. https://doi.org/10.1145/3313831.3376727.

Lou, Y., P. C. Abrami, J. C. Spence, C. Poulsen, B. Chambers, and S. d'Apollonia. 1996. "Within-Class Grouping: A Meta-Analysis." *Review of Educational Research* 66, no. 4: 423–458. https://doi.org/10.3102/00346543066004423.

Meyer, A., Y. Attali, M. Bar-Hillel, S. Frederick, and D. Kahneman. 2024. "Cognitive Reflection Is a Distinct and Measurable Trait." *Proceedings of the National Academy of Sciences of the United States of America* 121, no. 49: e2409191121. https://doi.org/10.1073/pnas.2409191121.

Molenaar, I., S. de Mooij, R. Azevedo, M. Bannert, S. Järvelä, and D. Gašević. 2023. "Measuring Self-Regulated Learning and the Role of AI: Five Years of Research Using Multimodal Multichannel Data." *Computers in Human Behavior* 139: 107540. https://doi.org/10.1016/j.chb.2022.107540.

Ng, D. T. K., J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. 2021. "Conceptualizing AI Literacy: An Exploratory Review." *Computers and Education: Artificial Intelligence* 2: 100041. https://doi.org/10.1016/j.caeai.2021.100041.

Otto, N., A. Böckers, T. Shiozawa, et al. 2024. "Profiling Learning Strategies of Medical Students: A Person-Centred Approach." *Medical Education* 58, no. 11: 1304–1314. https://doi.org/10.1111/medu.15388.

Parasuraman, R., and D. H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, no. 3: 381–410. https://doi.org/10.1177/0018720810376055.

Rosen, N. O., and B. Knäuper. 2009. "A Little Uncertainty Goes a Long Way: State and Trait Differences in Uncertainty Interact to Increase Information Seeking but Also Increase Worry." *Health Communication* 24, no. 3: 228–238. https://doi.org/10.1080/10410230902804125.

Şahin, M. G., and Y. Yıldırım. 2024. "The General Attitudes Towards Artificial Intelligence (GAAIS): A Meta-Analytic Reliability Generalisation Study." *International Journal of Assessment Tools in Education* 11, no. 2: 303–319. https://doi.org/10.21449/ijate.1369023.

Sailer, M., and L. Homner. 2020. "The Gamification of Learning: A Meta-Analysis." *Educational Psychology Review* 32, no. 1: 77–112. https://doi.org/10.1007/s10648-019-09498-w.

Sasson Lazovsky, G., T. Raz, and Y. N. Kenett. 2024. "The Art of Creative Inquiry—From Question Asking to Prompt Engineering." *Journal of Creative Behavior* 59, no. 1: e671. https://doi.org/10.1002/jocb.671.

Schepman, A., and P. Rodway. 2020. "Initial Validation of the General Attitudes Towards Artificial Intelligence Scale." *Computers in Human Behavior Reports* 1: 100014. https://doi.org/10.1016/j.chbr.2020.100014.

Schepman, A., and P. Rodway. 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, no. 13: 2724–2741. https://doi.org/10.1080/10447318.2022.2085400.

Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things." *American Journal of Psychology* 15, no. 1: 72–101. https://doi.org/10.2307/1412159.

Spurk, D., A. Hirschi, M. Wang, D. Valero, and S. Kauffeld. 2020. "Latent Profile Analysis: A Review and 'How to' Guide of Its Application Within Vocational Behaviour Research." *Journal of Vocational Behavior* 120: 103445. https://doi.org/10.1016/j.jvb.2020.103445.

Sun, L., and L. Zhou. 2024. "Does Generative Artificial Intelligence Improve the Academic Achievement of College Students? A Meta-Analysis." *Journal of Educational Computing Research* 62, no. 7: 1676–1713. https://doi.org/10.1177/07356331241277937.

Svedholm-Häkkinen, A. M. 2025. "Types of Analytic Thinkers." *Journal of Personality* 94, no. 2: 207–225. https://doi.org/10.1111/jopy.13025.

Täschner, J., T. Dicke, S. Reinhold, and D. Holzberger. 2025. ""Yes, I Can!" A Systematic Review and Meta-Analysis of Intervention Studies Promoting Teacher Self-Efficacy." *Review of Educational Research* 95, no. 1: 3–52. https://doi.org/10.3102/00346543231221499.

Thomson, K. S., and D. M. Oppenheimer. 2016. "Investigating an Alternate Form of the Cognitive Reflection Test." *Judgment and Decision Making* 11, no. 1: 99–113. https://doi.org/10.1017/S1930297500007622.

Toplak, M. E., R. F. West, and K. E. Stanovich. 2011. "The Cognitive Reflection Test as a Predictor of Performance on Heuristics-And-Biases Tasks." *Memory & Cognition* 39, no. 7: 1275–1289. https://doi.org/10.3758/s13421-011-0104-1.

Tylén, K., R. Fusaroli, S. M. Østergaard, P. Smith, and J. Arnoldi. 2023. "The Social Route to Abstraction: Interaction and Diversity Enhance Performance and Transfer in a Rule-Based Categorisation Task." *Cognitive Science* 47: e13338. https://doi.org/10.1111/cogs.13338.

Urban, M., C. Brom, J. Lukavský, et al. 2025. ""ChatGPT Can Make Mistakes. Check Important Info." Epistemic Beliefs and Metacognitive Accuracy in Students' Integration of ChatGPT Content Into Academic Writing." *British Journal of Educational Technology* 56, no. 5: 1897–1918. https://doi.org/10.1111/bjet.13591.

Weidlich, J., D. Gašević, H. Drachsler, and P. Kirschner. 2025. "ChatGPT in Education: An Effect in Search of a Cause." *Journal of Computer Assisted Learning* 41, no. 5: e70105. https://doi.org/10.1111/jcal.70105.

Wingerter, T. L., T. Straub, and S. Schweitzer. 2025. "Mitigating Automation Bias in Generative AI Through Nudges: A Cognitive Reflection Test Study." *Procedia Computer Science* 270: 2106–2114. https://doi.org/10.1016/j.procs.2025.09.331.

Xu, X., L. Qiao, N. Cheng, H. Liu, and W. Zhao. 2025. "Enhancing Self-Regulated Learning and Learning Experience in Generative AI Environments: The Critical Role of Metacognitive Support." *British Journal of Educational Technology* 56, no. 5: 1842–1863. https://doi.org/10.1111/bjet.13599.

Yang, M., S. Jiang, B. Li, et al. 2025. "Analysing Nontraditional Students' ChatGPT Interaction, Engagement, Self-Efficacy and Performance: A Mixed-Methods Approach." *British Journal of Educational Technology* 56, no. 5: 1973–2000. https://doi.org/10.1111/bjet.13588.

Yusuf, A., N. Pervin, M. Román-González, and N. Md Noor. 2024. "Generative AI in Education and Research: A Systematic Mapping Review." *Review of Education* 12, no. 2: e3489. https://doi.org/10.1002/rev3.3489.

Zhang, C., M. Hu, W. Wu, Y. Chen, F. Kamran, and X. Wang. 2025. "A Profile Analysis of Pre-Service Teachers' AI Acceptance: Combining Behavioral, Technological, and Human Factors." *Teaching and Teacher Education* 163: 105086. https://doi.org/10.1016/j.tate.2025.105086.

**Appendix A: Demographics Questionnaire.**

*Administration:* Pre-activity.

| Item | Variable | Question/response options |
|---|---|---|
| 1.1 | Gender | What is your gender? (Male/Female/Other/Prefer not to say) |
| 1.2 | Age | How old are you? (Open numeric response) |
| 1.3 | Year of Study | What year are you in? (Grade 1–3 for HS/Year 3–5 for UG/MSc Year 1–2 for PG) |
| 1.4 | Department/Field | What is your field/department? (General/Science Track/Computer Science/Engineering/etc.) |
| 1.5 | Education Level | What is your education level? (Upper Secondary/Undergraduate/Postgraduate) |

**Appendix B: GenAI Experience Questionnaire.**

*Administration:* Pre-activity.

*Scoring:* Experience Score = duration-of-use code (2.A1: 0 = never used, 1 = less than 1 month, 2 = 1–3 months, 3 = 4–6 months, 4 = 7–12 months, 5 = more than 1 year) + frequency-of-use code (2.A2: 0 = never, 1 = 2–3 times per month, 2 = once a week, 3 = 2–3 times per week, 4 = 4–6 times per week, 5 = daily). Possible range: 0–10.

| Item | Question |
|---|---|
| 2.A1 | How long have you been using GenAI tools? (Less than 1 month/1–3 months/4–6 months/7–12 months/More than 1 year) |
| 2.A2 | How often do you use GenAI tools? (Daily/4–6 times per week/2–3 times per week/Once a week/2–3 times per month) |
| 2.A3 | How would you rate your prompting capability? (1 = Not at all capable to 5 = Extremely capable) |
| 2.A4 | How much do you trust GenAI responses? (1 = Not at all to 5 = Completely) |
| 2.A5 | How often do you verify GenAI responses? (Always/Often/Sometimes/Rarely/Never) |

**Appendix C: Cognitive Reflection Test (CRT).**

*Administration:* Pre-activity|Scoring: 0 = intuitive/incorrect, 1 = reflective/correct|Total Score Range: 0–3.

*Purpose:* Measures tendency towards intuitive vs. analytical thinking.

| Item | Question | Intuitive | Correct |
|---|---|---|---|
| CRT_1 | A bat and a ball cost €1.10 in total. The bat costs €1.00 more than the ball. How much does the ball cost? | 10 cents | 5 cents |
| CRT_2 | If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? | 100 min | 5 min |
| CRT_3 | In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? | 24 days | 47 days |

**Appendix D: Intolerance of Uncertainty Scale—Short Form (IUS).**

*Administration:* Pre-activity|Scale: 1 = Not at all characteristic of me to 5 = Entirely characteristic of me|Total Score Range: 6–30.

*Purpose:* Measures dispositional intolerance of uncertainty.

| Item | Statement |
|---|---|
| IUS_1 | Unforeseen events upset me greatly. |
| IUS_2 | It frustrates me not having all the information I need. |
| IUS_3 | One should always look ahead so as to avoid surprises. |
| IUS_4 | A small unforeseen event can spoil everything, even with the best of planning. |
| IUS_5 | I always want to know what the future has in store for me. |
| IUS_6 | I can't stand being taken by surprise. |

**Appendix E: General Attitudes Towards Artificial Intelligence Scale (GAAIS) (see Tables E1 and E2).**

*Administration:* Pre-activity|Scale: 1 = Strongly Disagree to 5 = Strongly Agree|Score Range: 1–5 (mean) for each subscale.

**TABLE E1** | 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. |

**TABLE E2** | Negative attitudes subscale (negatively-worded items).

| 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 F: Technology Acceptance Model Scales (see Tables F1 and F2).**

*Administration:* Post-activity|Scale: 1 = Strongly Disagree to 7 = Strongly Agree|Score Range: 1–7 (mean) for each subscale.

**TABLE F1** | Perceived usefulness (PU).

| 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. |

**TABLE F2** | Perceived ease of use (PEU).

| 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 G: Prompting Self-Efficacy Scale (PSES).**

*Administration:* Post-activity|Scale: 1 = Not at all confident to 7 = Completely confident|Score Range: 1.0–7.0 (mean).

*Purpose:* Measures self-efficacy in AI prompting abilities.

*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? |

**Appendix H: Game Master Observation Rubric and Mystery Effectiveness Scoring.**

*Administration:* During activity (live GM observation); reasoning quality scored from each group's submitted solution.

The group's assigned Game Master completed a structured observation rubric during each session. 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 from the group's submitted solution and supporting argument, on a continuous 0–1 scale capturing logical coherence and evidence-based reasoning. In applying the anchors below, the Game Master drew on the problem-solving criteria recorded in the observation rubric (systematic analysis of the evidence, cross-referencing of evidence pieces and critical questioning of conclusions). Intermediate values (for example, 0.60) were permitted where a group's reasoning fell between two anchors.

| Score | Anchor description |
|---|---|
| 0.00 | No coherent reasoning. The conclusion is asserted without logical justification; available evidence is ignored, misread or contradicted. |
| 0.25 | 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. |
| 0.50 | 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. |
| 0.75 | 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. |

**Mystery Effectiveness Composition**

Observer-rated group performance measure (0–100, continuous). Group level: all members of a group share 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 I: Learning Behaviour With Emerging Technologies (LBET) Coding Scheme.**

*Application:* Researcher coding of the logged prompts. Adapted from Bloom's Digital Taxonomy.

| 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 |