

IRSTI: 14.01.85

AI UTILIZATION, COGNITIVE OFFLOADING, AND CRITICAL THINKING IN HIGHER CHEMICAL EDUCATION: A CROSS-SECTIONAL STUDY

Kamil R. Kabzhalelov^{*} , Zhanar K. Korganbayeva¹ , Marzhan A. Nurbekova² ¹Abai Kazakh National Pedagogical University, Almaty, Kazakhstan²Kazakh National Women's Teacher Training University^{*} Corresponding author's e-mail: kamil_kabzhalelov@outlook.com

ABSTRACT

The widespread adoption of generative artificial intelligence (AI) has intensified concerns about cognitive offloading and its relationship with critical thinking in higher education. This cross-sectional survey investigated the relationships among AI-related cognitive offloading, domain-specific AI use in chemistry education, and multidimensional critical thinking capacity among students and teaching staff predominantly from in Abai Kazakh National Pedagogical University (Almaty). Data were collected from 320 participants (80.3% female; 85.3% aged 17-25) during December 2025 – January 2026. Three instruments were administered: the AI Usage and Cognitive Offloading Scale (AICOS; author-developed; N = 320), the Critical Thinking Questionnaire (CThQ; Kobylarek et al., 2022; N = 320), and the AI in Chemistry Education Scale (AICES; author-developed; N = 305). AICOS and AICES are novel instruments that have not yet undergone independent external validation. Confirmatory factor analysis, reliability analysis, Welch's ANOVA with Games-Howell post-hoc tests, and bivariate correlation analyses were performed in Jamovi 2.5. The CThQ demonstrated acceptable structural validity (CFI = 939, RMSEA = 084) and excellent overall reliability ($\alpha = .916$). Most participants (59.1%) scored in the high normative range for critical thinking, whereas cognitive offloading was predominantly moderate (67.8%). No statistically significant correlations were found among AICOS, CThQ, and AICES (all $|r| < .05$, all $p > .40$). Age-group differences in AICOS scores were observed and remained statistically significant after adjustment for educational level, whereas gender did not predict any outcome variable. These findings should be interpreted cautiously given the ceiling effect in CThQ scores, the limited external validation of the author-developed scales, and the cross-sectional self-report design.

Keywords: Generative artificial intelligence; Critical thinking; Chemical education; Cognitive offloading; Higher education; Chemistry teacher education.

1 INTRODUCTION

The rapid proliferation of generative artificial intelligence (AI) tools has fundamentally altered the landscape of higher education within a remarkably short period. Since the public release of ChatGPT, GenAI adoption among university students has expanded rapidly (Chan & Hu, 2023). This technological shift has generated substantial scholarly debate concerning its educational consequences – particularly with regard to whether AI tool use enhances or undermines the higher-order cognitive skills that universities are designed to cultivate (Kasneci et al., 2023).

Among these human capacities, critical thinking stands as the most debated and theoretically central. Broadly understood as the disciplined, self-regulated process of conceptualising, applying, analysing, evaluating, and creating knowledge through active intellectual engagement (Facione, 1990), critical thinking is also recognised as a cornerstone competency for twenty-first-century life and work.

The primary theoretical mechanism through which AI use may influence critical thinking is cognitive offloading—the process by which individuals rely on external artefacts to reduce the cognitive load imposed on internal memory and reasoning systems (Risko & Gilbert, 2016).

According to this framework, humans strategically distribute cognitive labour between their own mental processes and external tools. Importantly, offloading is not inherently detrimental: when routine, low-level cognitive tasks are delegated to external systems, internal resources may be

reallocated to more complex operations. This adaptive strategy—referred to by Risko and Gilbert (2016) and later by Murphy (2023) as strategic offloading—suggests that the relationship between tool use and cognition is not zero-sum but potentially complementary.

The educational problem arises when offloading becomes non-selective – when students delegate not merely low cognitive tasks, but also analytical reasoning, evaluative judgement, and creative synthesis to AI systems. Sweller's (2011) cognitive load theory provides a complementary explanatory lens: the theory distinguishes between intrinsic cognitive load (the inherent complexity of a task), extraneous load (inefficient processing demands), and germane load (the productive mental effort that builds lasting cognitive schemas). When AI tools systematically absorb the intrinsic load of intellectually demanding tasks, learners are deprived of the productive cognitive struggle that schema formation and deep learning require.

The empirical literature on the relationship between AI tool use and critical thinking has grown rapidly since 2023, but its findings are notably inconsistent. Three broad positions can be identified.

The first position holds that AI dependence possibly erodes critical thinking through cognitive offloading mechanisms. Gerlich (2025), in a large-scale study, found a significant negative correlation between AI tool frequency of use and critical thinking scores, and argued that habitual AI reliance constitutes a form of cognitive offloading that progressively displaces independent analytical engagement. Corroborating evidence comes from Tian and Zhang (2025), whose moderated mediation study with 580 Chinese university students demonstrated a significant negative correlation between AI dependence and critical thinking, with cognitive fatigue partially mediating this relationship. Tian and Zhang's model revealed a dual-pathway mechanism: AI dependence leads to cognitive fatigue through both underload (insufficient intellectual stimulation when AI absorbs the thinking) and overload (the taxing demand of processing large volumes of AI-generated content), and this fatigue in turn impairs higher-order reasoning. Empirical support for the erosion position is further supplied by Yasmin Khairani Zakaria et al. (2025), who found that AI-dependent students demonstrated consistently lower engagement in deep analysis and critical evaluation of ill-structured problems.

The second position argues that AI use can augment rather than diminish critical thinking when the mode of engagement is reflective rather than passive. Mollick and Mollick (2023) found that structured AI-augmented activities – in which students were explicitly asked to critique, interrogate, and revise AI outputs – produced higher-quality analytical reasoning than unassisted control conditions. Yilmaz and Yilmaz (2023) similarly reported that prompting students to use ChatGPT as a Socratic interlocutor, rather than as an answer-generator, enhanced argumentative depth. Providing empirical support for this reflective paradigm, Yavich (2025) demonstrated in a controlled experiment that students engaged in pedagogically structured AI activities – accompanied by instructor guidance, formative feedback, and metacognitive tasks – achieved significantly higher critical thinking scores on the Watson – Glaser scale compared to traditional control conditions.

The third position emphasises population-specific and context-specific factors. Tian and Zhang (2025) demonstrated that information literacy – the capacity to locate, evaluate, and critically apply information – served as a boundary condition, buffering the negative direct effect of AI dependence on CT while paradoxically amplifying cognitive fatigue among heavy AI users with high literacy, due to the elevated monitoring demands that literate scrutiny of AI outputs entails. In a similar vein, Yavich (2025) emphasized that pedagogical scaffolding and instructor feedback serve as vital boundary conditions, determining whether AI adoption results in cognitive inertia or enhanced intellectual autonomy. These inconsistencies across the literature suggest that the relationship between AI use and critical thinking is not linear and unidirectional.

Existing studies on AI and cognitive offloading have been conducted almost exclusively in general academic or interdisciplinary settings, with low attention to disciplinary specificity. This is especially important because there is still a debate in the literature about whether critical thinking is a general skill or whether it is deeply tied to a specific area of activity (Tiruneh, De Cock and Elen, 2017). Chemistry represents a domain with a distinctive cognitive signature: it demands the

integration of symbolic-representational reasoning (chemical formulae, reaction mechanisms, stoichiometric calculations), spatial-conceptual modelling (molecular geometry, orbital hybridisation), and experimental data analysis (laboratory interpretation, error quantification).

The Critical Thinking Questionnaire (CThQ) developed by Kobylarek et al. (2022) offers a theoretically coherent and practically accessible alternative, operationalising critical thinking across the six levels of Bloom's (1956) revised taxonomy: Remembering, Understanding, Applying, Analysing, Evaluating, and Creating. This taxonomy-based approach permits researchers to identify not merely whether AI use is associated with lower overall CT scores, but which specific cognitive domains.

Synthesising the preceding review, interconnected gaps in the existing literature can be identified:

- first, despite a growing body of research on AI use and critical thinking, the authors found only a limited number of studies specifically addressing the relationship between artificial intelligence and critical thinking in the context of chemistry education.

- second, to date, the taxonomy-based CThQ developed by Kobylarek et al. (2022) does not appear to have been employed to examine AI-related cognitive offloading. As a result, it remains unclear which specific Bloom's-level cognitive domains might be differentially associated with various patterns of AI use.

- third, while Tian and Zhang (2025) and Gerlich (2025) provide important evidence of a negative AI–CT association in general and Chinese university populations, it is unknown whether these patterns replicate in Kazakhstani pedagogical university samples.

The present study aims to investigate the relationships among AI-related cognitive offloading, domain-specific AI use in chemistry education, and multidimensional critical thinking capacity – as measured by the CThQ (Kobylarek et al., 2022). The study additionally examines demographic moderators (age, gender) and evaluates the structural validity of the CThQ in this new cultural context. By doing so, it seeks to extend the theoretical framework established by Gerlich (2025) and Tian and Zhang (2025) into a chemistry-specific, pedagogical university, Central Asian context – thereby addressing all identified gaps simultaneously.

RQ1. What are the levels of cognitive offloading via AI tools (AICOS), critical thinking (CThQ), and domain-specific AI use in chemistry education (AICES) among students and faculty at Kazakhstani pedagogical universities, and how are participants distributed across low, average, and high performance levels?

RQ2. Are there statistically significant associations between cognitive offloading (AICOS), critical thinking (CThQ), and AI use in chemistry education (AICES)?

RQ3. Do age, gender significantly predict cognitive offloading (AICOS) and critical thinking (CThQ) scores?

RQ4. Does the six-factor structure of the CThQ (Kobylarek et al., 2022) demonstrate adequate confirmatory factor analytic fit in a Kazakhstani chemistry education sample?

Hypotheses of this study:

H1. Higher levels of cognitive offloading (AICOS) will be significantly and negatively associated with overall critical thinking scores (CThQ total), consistent with cognitive load theory (Sweller, 2011), Gerlich (2025), and Tian and Zhang (2025).

H2. Higher levels of domain-specific AI use in chemistry (AICES) will be negatively associated with the CThQ.

H3. Younger participants (age 17–25) will demonstrate higher cognitive offloading scores than older participants.

H4. Gender will not significantly predict cognitive offloading or critical thinking scores.

The present study contributes:

At the first, the study provides empirical data regarding the relationships among AI-related cognitive offloading, domain-specific AI use in chemistry education, and critical thinking within a Kazakhstani higher education context;

Secondly, the study introduces two author-developed instruments (AICOS and AICES) designed to assess AI-related cognitive offloading and AI engagement in chemistry education. Their psychometric properties are evaluated as an initial step in instrument development;

Thirdly, the study provides an initial psychometric evaluation of the Russian and Kazakh translations of the CThQ in a Kazakhstani sample;

Finally, the findings provide evidence that may inform the responsible integration of generative AI into chemistry teacher education and support future research on AI literacy and critical thinking.

2 MATERIALS AND METHODS

2.1 Research designs

This study adopted a cross-sectional quantitative survey design to examine the relationships among AI-related cognitive offloading, domain-specific AI use in chemistry education, and critical thinking capacity.

A convenience sampling strategy with voluntary participation was employed. A total of 320 participants (mainly from Abai Kazakh National Pedagogical University (KazNPU) , additional participants were recruited from several other educational institutions in Kazakhstan, including Kazakh National Women's Teacher Training University, Al-Farabi Kazakh National University, SDU University, Ualikhanov University) completed the AI Usage and Cognitive Offloading Scale (AICOS) and the Critical Thinking Questionnaire (CThQ). Of these, 305 participants also completed the AI in Chemistry Education Scale (AICES). The study was approved by the Ethics Committee of the Abai Kazakh National Pedagogical University. A detailed summary of participants' demographic characteristics is presented in table 1.

Table 1. Demographic Characteristics of the study

Characteristic	Category	N (AICOS)	%	N (CThQ)	%	N (AICES)	%
Gender	Female	257	80.3	257	80.3	229	75.1
	Male	55	17.2	55	17.2	70	22.9
	Prefer not to say	8	2.5	8	2.5	6	2.0
Age	17–25	273	85.3	273	85.3	297	97.4
	26–35	17	5.3	17	5.3	2	0.7
	36–45	9	2.8	9	2.8	1	0.3
	46–55	12	3.8	12	3.8	4	1.3
	56+	9	2.8	9	2.8	1	0.3
Education level	Undergraduate	252	78.8	252	78.8	286	93.8
	Master's	30	9.4	30	9.4	14	4.6
	Doctoral (PhD) and above	38	11.9	38	11.9	5	1.6

2.2 AI tools for research

AI Usage and Cognitive Offloading Scale (AICOS; Appendix A). The AICOS was developed by the research team as no existing validated instrument measured AI-related cognitive offloading in higher education at the time of data collection. The development procedure proceeded in three stages. First, the theoretical framework of Risko and Gilbert (2016) and Sweller's (2011) cognitive load theory were used to identify four dimensions: frequency and context of AI use, delegation of academic cognitive effort, impact on autonomous motivation, and trust in AI outputs. Second, an initial pool of 34 items was generated by the first author, covering all four dimensions. Third, two co-authors independently reviewed each item for content representativeness, clarity, and absence of double-barrelled phrasing; 13 items were revised or removed based on consensus. Finally, 21-item scale was used in this study. No external expert panel or test-retest reliability study was conducted prior to the main data collection. Items use a six-point Likert scale (1 = Strongly Disagree, 6 =

Strongly Agree); seven items are reverse-coded (items 6, 11, 13, 15, 16, 18, 20), so that higher total scores consistently reflect greater cognitive offloading (range: 21–126).

Critical Thinking Questionnaire (CThQ; Kobylarek et al., 2022). The CThQ is a validated 25-item self-report instrument developed at the University of Wrocław (Poland) and grounded in Bloom's (1956) taxonomy of educational objectives. It assesses critical thinking across six hierarchically ordered subscales: Remembering (3 items), Understanding (4 items), Applying (4 items), Analysing (4 items), Evaluating (4 items), and Creating (6 items). Items are rated on a five-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree), with four items requiring reverse scoring (items 4, 9, 19, 22). Raw scores are summed for each subscale and for the total scale (range: 25–125). The instrument demonstrated satisfactory psychometric properties in its original Polish standardisation study (overall $\alpha = .877$, $N = 198$; Kobylarek et al., 2022). Normative cut-scores are: Low (25–58), Average (59–92), and High (93–125). The CThQ was translated into Russian and Kazakh using a forward-translation procedure. Because back translation and formal cross-cultural adaptation procedures were not conducted, the present study should be regarded as an initial psychometric evaluation.

AI in Chemistry Education Scale (AICES; Appendix B). The AICES was developed to capture domain-specific patterns of AI use within chemistry learning and teaching. The development procedure mirrored that of the AICOS: theoretical item generation based on the main cognitive activities in university chemistry courses (conceptual understanding, laboratory work, quantitative problem-solving, academic writing, and critical evaluation of AI outputs), followed by co-author content review and reduction from 12 candidate items to 7. Items use a six-point Likert scale (1 = Strongly Disagree, 6 = Strongly Agree); item 4 requires reverse coding. Total scores range from 7 to 42.

Data were collected during December–January 2026 via three separate Google Forms surveys distributed to participants through university faculty coordinators and group academic representatives. Participation was entirely voluntary and anonymous; no personal identifying information was collected. Respondents were informed of the study's purpose, their right to withdraw at any time without consequence, and the exclusively academic use of their responses. Survey links were shared through university internal communication channels and chemistry department group chats. All three surveys were completed independently by each participant in a self-paced online format. To minimize duplicate submissions, the Google Forms setting limiting each Google account to a single response was enabled throughout the data collection period.

2.3 Data analysis

All statistical analyses were conducted using Jamovi (version 2.5.6; The jamovi project, 2024), an open-source statistical software platform. The following analytical procedures were applied in sequence:

Descriptive statistics (mean, median, standard deviation, range, skewness, kurtosis) were calculated for all three scale total scores and for each CThQ subscale. Normality was assessed via the Shapiro-Wilk test, with supplementary evaluation of skewness (acceptable range: ± 2) and kurtosis (acceptable range: ± 7) following Hair et al. (2019).

Participant-level classification for the AICOS and AICES was performed using the $M \pm 1$ SD rule: participants scoring below $M - 1$ SD were classified into the lower distribution group, those scoring above $M + 1$ SD into the upper distribution group, and the remaining participants into the central distribution group (Tabachnick & Fidell, 2019). For the CThQ, Kobylarek et al.'s (2022) original normative cut-scores were applied.

Internal consistency reliability was estimated using Cronbach's α and McDonald's ω for the total AICOS scale and for each of the six CThQ subscales. The conventional threshold of $\alpha \geq .70$ was applied (Nunnally & Bernstein, 1994; Taber, 2018).

Confirmatory factor analysis (CFA) was conducted for the CThQ using maximum likelihood estimation, specifying the six-factor structure proposed by Kobylarek et al. (2022). Model fit was

evaluated using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) with 90% confidence interval, and the Standardised Root Mean Square Residual (SRMR). Acceptable fit was defined as CFI/TLI $\geq$.90 and RMSEA $\leq$.08 (Hu & Bentler, 1999; Kline, 2015).

Bivariate correlations between the three scale total scores were calculated using both Pearson product-moment and Spearman rank-order coefficients, given the mixed normality results across scales. Statistical significance was set at $\alpha = .05$ (two-tailed).

Group difference analyses used Welch's one-way ANOVA (robust to unequal group sizes and heterogeneity of variance) to compare AICOS, CThQ, and AICES scores across age groups, with Tukey post-hoc tests for pairwise comparisons. The independent-samples Welch's t-test was applied for binary gender comparisons. A chi-square test of independence examined CThQ level category distributions by gender. Effect sizes were calculated as η^2 for ANOVA and r for pairwise comparisons.

3 RESULTS AND DISCUSSION

3.1 Descriptive statistics and participant-level classification

A total of 320 participants completed the AI Usage and Cognitive Offloading Scale (AICOS) and the Critical Thinking Questionnaire (CThQ; Kobylarek et al., 2022), while 305 participants completed the AI in Chemistry Education Scale (AICES). All three instruments were administered concurrently via Google Forms during a single data-collection period from December to January 2026.

Descriptive statistics for all three scales are presented in Table 2. The AICOS total score yielded a mean of 66.5 (SD = 13.0), with observed scores ranging from 26 to 105. A Shapiro-Wilk test indicated that the AICOS distribution did not significantly deviate from normality ($W = 0.994$, $p = .282$; skewness = -0.18 , kurtosis = -0.10), permitting the subsequent use of parametric statistical procedures for this scale. The CThQ overall score ($M = 95.8$, $SD = 15.5$) and the AICES score ($M = 23.9$, $SD = 8.26$) showed statistically non-normal distributions (both $p < .001$); however, skewness and kurtosis values remained within acceptable limits ($|skewness| < 1.0$, $kurtosis < 1.0$), indicating near-symmetrical distributions with no severe departures from normality (Hair et al., 2019).

Table 2. Descriptive Statistics for All Three Instruments

Scale	N	Mean (SD)	Range	SW p
AICOS (Cognitive Offloading)	320	66.5 (13.0)	26–105	.282
CThQ (Critical Thinking)	320	95.8 (15.5)	43–125	< .001
AICES (AI in Chemistry)	305	23.9 (8.3)	7–42	< .001

*Note: SW p = Shapiro-Wilk p-value; SD = Standard Deviation; AICOS = AI Usage and Cognitive Offloading Scale; CThQ = Critical Thinking Questionnaire (Kobylarek et al., 2022); AICES = AI in Chemistry Education Scale.

To enable interpretable group comparisons for the researcher-developed instruments (AICOS and AICES), participants were classified into three proficiency levels using a mean $\pm$ one standard deviation (see table 3). This approach is standard practice when normative data for custom-developed scales are unavailable (Tabachnick & Fidell, 2019). For the CThQ, the original normative cut-scores proposed by Kobylarek et al. (2022) were applied: Low (25–58), Average (59–92), and High (93–125).

Applying this relative classification, the majority of participants (67.8%, $n = 217$) fell into the central distribution group for AI-driven cognitive offloading (AICOS), with 15.0% ($n = 48$) categorized within the lower distribution group and 17.2% ($n = 55$) within the upper distribution group. The AICOS cut-off thresholds were calculated at < 53.5 for the lower group and > 79.5 for the upper group. For the CThQ, only 3 respondents (0.9%) scored within the lower score tier, whereas 189 (59.1%) were situated in the upper score tier and 128 (40.0%) fell into the central score tier. To

characterise the degree of score concentration more precisely, 34 participants (10.6%) scored at or above 113 – the upper 10% of the 25–125 scale range – and 3 participants (0.9%) achieved the maximum possible score of 125. Taken together, 11.6% of the sample ($n = 37$) scored in the uppermost portion of the scale, indicating a pronounced ceiling effect that severely restricts score variance and constitutes the primary statistical explanation for the null correlations observed. This distributional pattern reflects a positive skew toward the upper score tier in this specific sample, which may align with the elevated educational background of participants (predominantly undergraduate, master's, and doctoral students alongside teaching staff).

Table 3. Level Classification Summary for Each Scale

Scale	Classification Method	Low threshold	High threshold	n Lower distribution group (%)	n Upper distribution group (%)	n Central distribution group (%)
AICOS	$M \pm 1SD$	< 53.5	> 79.5	48 (15.0%)	55 (17.2%)	217 (67.8%)
CThQ	Kobylarek norms	< 59.0	> 92.0	3 (0.9%)	189 (59.1%)	128 (40%)
AICES	$M \pm 1SD$	< 15.6	> 32.2	49 (16%)	43 (14%)	213 (70%)

**Note: AICOS and AICES thresholds based on $M \pm 1SD$ rule. CThQ thresholds based on Kobylarek et al. (2022).*

3.2 Reliability Analysis

3.2.1 Exploratory factor analysis of AICES

Because AICES was developed as a new instrument, an exploratory factor analysis (EFA) was conducted to examine its underlying structure. The analysis demonstrated good sampling adequacy ($KMO = 0.843$), and Bartlett's test of sphericity was significant, $\chi^2(21) = 910$, $p < 0.001$, indicating that the data were suitable for factor analysis. Using minimum residual extraction with oblimin rotation, a two-factor solution was identified (table 4). However, several items demonstrated moderate cross-loadings across both factors, suggesting that AICES is not strictly unidimensional but rather reflects overlapping dimensions of AI engagement in chemistry learning. The two factors were moderately correlated ($r = 0.464$), indicating that they represent related rather than independent aspects of AI use.

Table 4. Factor loadings

Item	F1	F2
1	0.378	0.494
2	0.546	0.364
3	0.798	-
4	-	0.723
5	0.713	-
6	0.849	-
7	0.340	0.468

Given the cross-loading pattern, the moderate correlation between factors, and the exploratory nature of the present validation, AICES was retained as a single composite measure in the subsequent analyses. The identified multidimensionality should be examined further in future validation studies using larger and more diverse samples.

To evaluate the internal structure of the AICOS, an exploratory factor analysis (EFA) was conducted using minimum residuals extraction with oblimin rotation, which is appropriate for correlated factor structures ($N = 320$). Sampling adequacy was excellent ($KMO = 0.919$; Bartlett's $\chi^2(210) = 3785$, $p < 0.001$). The three-factor solution demonstrated acceptable model fit ($RMSEA = 0.068$ [90% CI: 0.059–0.077], $TLI = 0.913$) and explained 55.1% of total variance.

Factor 1 – AI Cognitive Dependency (10 items; $SS = 5.21$; 24.8% of variance) comprises items reflecting substitution of AI for memory, autonomous analysis, and independent motivation (e.g.,

AICOS-8: "I rely on AI to find and process information instead of remembering it myself"; AICOS-17: "I trust AI recommendations even when I do not understand how they were generated").

Factor 2 – Critical-Autonomous Engagement (7 items; SS = 3.99; 19.0%) comprises all reverse-coded items reflecting deliberate verification, critical scrutiny of AI outputs, and preference for independent problem-solving (e.g., AICOS-16: "I reflect on possible biases and errors in AI-generated responses"; AICOS-20: "I verify AI-generated information by comparing it with other sources").

Factor 3 – Frequency of AI Use (4 items; SS = 2.37; 11.3%) reflects the general daily regularity of AI tool use across educational and professional tasks (AICOS-1, AICOS-2, AICOS-3).

Inter-factor correlations were moderate: Dependency and Frequency correlated positively ($r = 0.463$), while Critical-Autonomous Engagement correlated negatively with both Dependency ($r = -0.237$) and Frequency ($r = -0.391$). Internal consistency was computed separately for each EFA-derived subscale. AI Cognitive Dependency (10 items) yielded $\alpha = 0.91$, $\omega = 0.91$; Critical-Autonomous Engagement (7 items) $\alpha = 0.88$, $\omega = 0.88$; Frequency of AI Use (4 items) $\alpha = 0.82$, $\omega = .83$. All three subscales exceeded the $\alpha \geq .70$ threshold (Taber, 2018). Internal consistency of the AICOS was satisfactory at the overall scale level (Cronbach's $\alpha = .728$), marginally exceeding the widely accepted threshold of 0.70 (Nunnally & Bernstein, 1994; Taber, 2018).

The CThQ demonstrated excellent internal consistency at the global level ($\alpha = 0.916$, McDonald's $\omega = 0.931$), substantially exceeding the original reliability reported by Kobylarek et al. (2022; $\alpha = 0.877$). Subscale-level reliabilities, compared with original values, are shown in Table 5.

Table 5. CThQ Subscale Reliability: Original vs. Present Study

Subscale (Bloom's domain)	Items	Original α	Present α	Present ω
Evaluating	4	0.712	0.794	0.803
Applying	4	0.644	0.776	0.777
Creating	6	0.554	0.692	0.756
Analysing	4	0.568	0.521	0.625
Understanding	4	0.574	0.488	0.600
Remembering	3	0.608	0.261	0.457
Overall CThQ	25	0.877	0.916	0.931

The Evaluating subscale showed the strongest reliability ($\alpha = 0.794$, $\omega = 0.803$), while the Remembering subscale showed markedly low reliability ($\alpha = 0.261$, $\omega = 0.457$). This finding requires methodological attention: the Remembering subscale contains item 19R(r) – a negatively worded item requiring reverse coding – and its low alpha coefficient suggests that the item may function differently in a Kazakhstani pedagogical context compared with the original Polish sample. The Analysing ($\alpha = 0.521$) and Understanding ($\alpha = 0.488$) subscales also fell below the 0.70 threshold.

3.2.2 Confirmatory Factor Analysis of the CThQ

To evaluate the structural validity of the CThQ in the present sample, a confirmatory factor analysis (CFA) was conducted using maximum likelihood estimation in Jamovi (version 2.5), specifying the six-factor structure proposed by Kobylarek et al. (2022): Evaluating, creating, analysing, remembering, understanding, and applying. Factor fit indices are presented in Table 6.

Table 6. Confirmatory Factor Analysis Fit Indices for the Six-Factor CThQ Model

CFI	TLI	RMSEA [90% CI]	χ^2 (df), p	SRMR
0.939	0.919	0.084 [0.067–0.102]	111(34), < 0.001	0.043

*Note: CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation with 90% CI; SRMR = Standardised Root Mean Square Residual. Recommended thresholds:

CFI/TLI ≥ 0.95 (excellent), ≥ 0.90 (acceptable); RMSEA ≤ 0.06 (excellent), ≤ 0.08 (acceptable) (Hu & Bentler, 1999; Kline, 2015).

The CFA results indicate acceptable, though not excellent, model fit: CFI = 0.939, TLI = 0.919, RMSEA = .084 [90% CI: .067–.102]. The CFI and TLI values exceed the .90 acceptability threshold but fall slightly below the .95 excellent-fit benchmark. The RMSEA is marginally above the strict .08 criteria but within the .10 maximum tolerance suggested by MacCallum et al. (1996). The statistically significant chi-square ($\chi^2 = 111$, $df = 34$, $p < .001$) is expected given the large sample size, as this test is sensitive to sample size and is not recommended as a stand-alone fit criterion (Kline, 2015).

Inspection of individual factor loadings revealed that all items loaded significantly on their hypothesised factors (all $p < .001$), with one notable exception: item 9C(r) – 'Everything already exists, so nothing completely new can be created' – yielded a non-significant and near-zero loading on the Creating factor ($\beta = -0.019$, $p = 0.799$). This item is reverse-scored, and its failure to load appropriately reinforces the hypothesis of item-level coding issues noted above. The high covariance between the Evaluating and Creating factors ($r = 0.965$, $p < 0.001$) also warrants attention and may indicate that these two constructs are not sufficiently distinct within this sample, possibly reflecting a general higher-order critical thinking disposition (Facione, 1990). Overall, the CFA provides provisional support for the six-factor structure of the CThQ in a Kazakhstani chemistry education context, though the fit is less than optimal.

3.2.3 Group differences by age and gender

Table 7 presents descriptive statistics, including 95% confidence intervals, for all three scales stratified by age group.

Table 7. Descriptive statistics by age group across all three scales

Age group	n	AICOS M (SD)	95% CI (for AICOS)	CThQ M (SD)	AICES M (SD)	AICOS Distribution Group
17–25	273	67.3 (13.0)	65.7–68.8	95.1 (14.9)	23.9 (8.3)	Central distribution
26–35	17	67.5 (9.7)	62.5–72.5	101.0 (13.0)	25.0 (2.8)	Central distribution
36–45	9	55.0 (9.4)*	47.8–62.2	106.4 (14.9)	–	Lower distribution
46–55	12	59.1 (13.0)	50.8–67.3	92.2 (26.0)	–	Lower distribution
56+	9	60.8 (13.3)	50.5–71.0	102.3 (13.6)	–	Lower distribution

*Note: AICOS: $M \pm 1SD$ classification thresholds: Low < 53.5 , High > 79.5 . * Significantly different from the 17–25 and 26–35 age groups according to Games–Howell post-hoc comparisons ($p = 0.027$ and $p = 0.039$, respectively). CThQ norms: Low < 59 , Average 59–92, High > 92 . AICES cells with insufficient n per group are marked.

A one-way Welch's ANOVA indicated a statistically significant effect of age on AICOS scores, $F(4, 21.70) = 4.57$, $p = 0.008$. The corresponding effect size was small ($\omega^2 = 0.031$). Games–Howell post-hoc comparisons showed that participants aged 36–45 years reported significantly lower AICOS scores than both the 17–25 age group (mean difference = 12.30, $p = 0.027$) and the 26–35 age group (mean difference = 12.50, $p = 0.039$). No other pairwise comparisons reached statistical significance. The 95% confidence intervals of the group means are presented in Table 7. Given the relatively small number of participants in the 36–45 age group ($n = 9$), these findings should be interpreted cautiously.

No significant age-group differences were found for the AICES total score (Welch's $F = 2.33$, $df_1 = 4$, $df_2 = 21.2$, $p = .090$), and gender did not significantly predict AICOS scores (all post-hoc $p > .28$) nor AICES scores (Welch's $F = 0.221$, $df_1 = 2$, $df_2 = 13.1$, $p = .805$). A chi-square test of independence for CThQ level (Low/Average/High) by gender was likewise non-significant ($\chi^2 = 2.84$,

df = 4, $p = .586$), confirming that gender did not moderate critical thinking level distribution in this sample.

3.2.4 Additional analysis of age-group differences

Because age group in the present sample was associated with participants' educational level, an additional multiple linear regression analysis (Cohen et al., 2013) was conducted to examine whether the observed age differences in AICOS scores remained after controlling for educational level. Age group and educational level were entered simultaneously as categorical predictors, with the 17–25 age group and undergraduate students serving as the reference categories.

The overall regression model was statistically significant, $F(7, 312) = 3.30$, $p = 0.002$, explaining 6.9% of the variance in AICOS scores ($R^2 = 0.069$; adjusted $R^2 = 0.048$, see table 8). After adjustment for educational level, participants aged 36–45 years, 46–55 years, and 56 years and older continued to report significantly lower AICOS scores than participants aged 17–25 years. Educational level was also independently associated with AICOS scores, with PhD participants and master's students reporting significantly higher scores than undergraduate students. These findings indicate that the observed age differences cannot be explained solely by differences in educational level within the sample.

Table 8. Multiple linear regression predicting AICOS scores from age group and educational level

Predictor	B	SE	t	p	95% CI
Intercept	65.52	1.07	61.23	< 0.001	63.42 to 67.63
Educational levels					
Master's vs Undergraduate	5.81	2.89	2.01	0.045	0.12 to 11.49
PhD or higher vs Undergraduate	10.77	4.60	2.34	0.020	1.71 to 19.82
Age groups					
26–35 vs 17–25	−6.78	4.49	−1.51	0.132	−15.61 to 2.05
36–45 vs 17–25	−20.74	5.91	−3.51	< 0.001	−32.37 to −9.11
46–55 vs 17–25	−16.38	5.39	−3.04	0.003	−26.99 to −5.77
56+ vs 17–25	−14.32	5.82	−2.46	0.014	−25.77 to −2.87

*Model statistics: $R^2 = 0.069$; Adjusted $R^2 = 0.048$; $F(7, 312) = 3.30$, $p = 0.002$.

3.2.5 Correlations between scales

The central research question of this study concerned the relationships between AI-related cognitive offloading, domain-specific AI use in chemistry education, and critical thinking capacity.

Table 9. Correlation Matrix: AICOS, CThQ, and AICES Total Scores (N = 303–320)

Variable	AICOS	CThQ	AICES
1. AICOS (Cognitive Offloading)	–	–	–
2. CThQ (Critical Thinking)	$r = 0.046$, $p = 0.417$ $\rho = 0.004$, $p = 0.944$	–	–
3. AICES (AI in Chemistry)	$r = 0.039$, $p = 0.493$	$r = -0.004$, $p = 0.945$	–

*Note. Pearson r reported above Spearman ρ for AICOS–CThQ pair. $df = N - 2$. Two-tailed significance testing. No correlations were statistically significant at the $\alpha = 0.05$ level.

Pearson product-moment and Spearman rank-order correlation coefficients between the three scale total scores are reported in table 9.

The correlation between AICOS and CThQ was negligible and non-significant in both Pearson ($r = 0.046$, $df = 318$, $p = 0.417$) and Spearman analyses ($\rho = 0.004$, $df = 318$, $p = 0.944$). Similarly, AICOS and AICES were not significantly correlated ($r = 0.039$, $df = 303$, $p = 0.493$), and neither

were CThQ and AICES ($r = -.004$, $df = 303$, $p = 0.945$). None of the three inter-scale correlations reached statistical significance, and all effect sizes were negligible ($|r| < .05$).

3.3 Critical thinking profile: A uniformly high-performing sample

The most notable CThQ-related finding is the score distribution: 59.1% of participants were classified as High-level critical thinkers under Kobylarek et al.'s (2022) normative framework, with only 0.9% in the Low category – a pattern consistent with positive educational selection in a sample comprising pedagogical university students, doctoral candidates, and chemistry lecturers. From the perspective of Bloom's (1956) taxonomy, the Evaluating ($\alpha = .794$) and Applying ($\alpha = .776$) subscales demonstrated both the strongest psychometric properties and the highest CFA loadings, suggesting these higher-order domains are reliably measured and prominent in this population. By contrast, the Remembering subscale yielded critically low reliability ($\alpha = .261$, vs. $\alpha = .608$ in Kobylarek et al., 2022), likely because the conceptual relevance of school-based memory recall differs for mature university students and working academics. Future studies should examine item 19R(r) for differential item functioning across educational levels.

3.3.1. Cognitive offloading and AI use in chemistry at average-level patterns

The AICOS mean of 66.5 ($SD = 13.0$) placed 67.8% of participants in the Average band, suggesting moderate rather than habitual AI reliance – a pattern consistent with the selective offloading described by Risko and Gilbert (2016), wherein cognitively competent individuals distribute cognitive labour contextually rather than uniformly. The significant age-related difference – with the 36–45 cohort reporting lower offloading ($M = 55.0$) than the 17–25 cohort ($M = 67.3$, $p = 0.039$) – likely reflects that older participants completed their foundational training before AI tools became pervasive, though this effect is substantially confounded with professional status and should not be attributed to age alone. The AICES mean of 23.9 ($SD = 8.26$) positioned respondents near the midpoint of the 7–42 scale, indicating moderate domain-specific AI adoption in chemistry. No significant gender effects were detected for any outcome variable, consistent with Strzelecki (2023), who found that gender does not moderate AI tool acceptance in higher education.

3.3.2 The null correlation in interpreting the absence of a relationship

No statistically significant correlations were observed among AICOS, CThQ, and AICES (all $|r| < 0.05$, all $p > 0.40$). Before attempting any substantive interpretation, three methodological conditions of the present study must be considered, as each is independently capable of producing null correlations regardless of any true underlying relationship.

The most consequential is the ceiling effect in CThQ scores. With 59.1% of participants in the High normative range and 20.6% scoring above 111, score variance was severely restricted. Restriction of range is a well-established statistical condition that deflates observed correlations toward zero even when a genuine relationship exists in more heterogeneous populations. Second, the AICOS total-scale reliability of $\alpha = .728$ is only marginally acceptable, and the CThQ showed below-threshold subscale reliabilities for three of six scales. Measurement error in both instruments could further attenuate true correlations. Third, the AICES composite score amalgamates conceptually distinct behaviours – frequency of use, cognitive dependency, uncritical trust, and perceived benefit – which may have opposing relationships with critical thinking that cancel each other when summed into a single score.

Given these three conditions, the null correlations cannot be interpreted as evidence that AI-related cognitive offloading and critical thinking are genuinely unrelated in this population. A study designed to avoid these limitations – employing a more heterogeneous sample, externally validated instruments, and separate AICES subscales – might produce substantially different results. Accordingly, H1 and H2 are not supported by the present data, but the substantive hypothesis that cognitive offloading negatively predicts critical thinking in chemistry education contexts remains open.

It is worth noting that the null finding is not consistent with prior research using different instruments and samples: Gerlich (2025) and Tian and Zhang (2025) both documented significant negative associations between AI dependence and critical thinking in more heterogeneous general-population samples. The divergence between those findings and the present null result may be explained by the methodological factors outlined above, rather than by a genuine population difference.

If the methodological explanations are nonetheless set aside and a conditional substantive interpretation is entertained, two possibilities may be considered. First, the present sample – predominantly enrolled in pedagogical chemistry programmes and including doctoral candidates and instructors – may engage in more selective AI use than the heterogeneous general-population samples studied by Gerlich (2025) and Tian and Zhang (2025). If the majority of participants offload primarily low-level tasks while retaining higher-order processing, cognitive offloading and critical thinking may not conflict in this specific subpopulation. Second, the high baseline critical thinking capacity of the sample may mean that AI use has not yet produced measurable effects. Both interpretations should be treated as directions for future investigation rather than conclusions of this study.

3.3.3 Age as a Moderating Variables including experience, expertise and cognitive habits

The age-graded AICOS pattern – older participants reporting lower offloading – offers a pedagogical nuance: experienced educators may model strategic AI use, embedding critical evaluation of AI outputs into disciplinary practice. Conversely, the 17–25 age group combined high AI cognitive offloading ($M = 67.3$) with high critical thinking scores (CTHQ $M = 95.1$), suggesting capable critical thinkers who nonetheless rely heavily on AI. This mirrors what Sweller (2011) frames as cognitive load externalisation risk: students may possess evaluative capacity without habitually engaging it – a latent critical thinking that remains dormant rather than deployed. The conditions activating this latency are a target for the study's planned qualitative interviews, which probe self-perception versus actual behaviour.

3.3.4 Implications for chemistry education and pedagogical practice

These findings carry three practical implications. Firstly, the high prevalence of strong critical thinking scores suggests the sample has the cognitive prerequisites to engage productively with AI; AI-augmented lab-report tasks requiring students to evaluate and critique AI-generated explanations against experimental data could activate higher-order Bloom's domains rather than passive Remembering-level delegation. Secondly, older academics' lower offloading positions experienced instructors to explicitly model selective AI use – retrieval and preliminary organisation via AI, while retaining ownership of analysis and argumentation. Third, in Kazakhstan's pedagogical-university context, future chemistry teachers who use AI uncritically risk transmitting that pattern to school-age learners. Curriculum policy should therefore address AI literacy explicitly within teacher education, embedding reflective AI use as a pedagogical competence (Holmes & Miao, 2023).

3.4 Challenges and future perspectives

Several limitations must be acknowledged. First, all three instruments rely on self-report, which is susceptible to social desirability bias. Future studies should triangulate survey data with behavioural measures such as AI interaction logs or performance-based CT assessments. Second, the cross-sectional design precludes causal inference: whether AI use affects critical thinking, or whether individuals with certain CT profiles self-select into particular AI use patterns, cannot be determined from the present data. Third, the sample is a convenience sample, is predominantly female (80.3%) and young (85.3% aged 17–25), and is drawn from a pedagogical specialisation. Observed age-group differences in AICOS may be closely related to professional status, and the findings are not generalisable to non-pedagogical, more gender-balanced, or more age-diverse higher education populations. Fourth, the AICOS and AICES were custom-developed instruments without prior external validation, and the AICOS reliability ($\alpha = 0.728$) is only marginally acceptable. Fifth,

potential item-level reverse-coding issues in the CThQ (particularly items 9C(r), 19R(r), and 22U(r)) may have systematically attenuated subscale reliability estimates, and future analyses should verify correct reverse coding before replication.

Future research should focus on this work in several ways. Longitudinal studies could assess whether AI tool adoption across an academic year predicts changes in CT scores after accounting for initial levels. Experimental or quasi-experimental designs—such as comparing structured AI-use training with unguided AI access among chemistry students—would afford stronger causal inferences. Mediation models featuring metacognitive awareness as a moderator may elucidate why the relationship between CT and AI offloading varies across populations. Lastly, qualitative data collection and analysis would provide valuable contextual information to complement the quantitative null results and to explore the behavioural factors shaping students' AI-use decisions in chemistry education.

4 CONCLUSION

This study investigated the relationships among AI-related cognitive offloading, domain-specific AI use in chemistry education, and multidimensional critical thinking among students and faculty from Kazakhstani higher education institutions. Participants generally demonstrated high critical thinking scores, whereas AI-related cognitive offloading and domain-specific AI use were predominantly concentrated within the central distribution of the sample.

No statistically significant relationships were identified among cognitive offloading, AI use in chemistry education, and critical thinking. Age-related differences were observed, participants aged 36-45 reported lower AICOS scores than the youngest age group; however, this result should be interpreted cautiously because of the relatively small sample size of the older age group. No significant gender differences were found. These findings suggest that patterns of AI use may vary across demographic groups, although no measurable association with critical thinking was observed in the present sample.

The psychometric evaluation provided acceptable, though not conclusive, support for the six-factor structure of the CThQ in the Kazakhstani sample. Although the overall reliability of the instrument was high, the low internal consistency of the Remembering subscale and the use of a forward-translation procedure without back translation indicate that the translated instrument requires further validation before broader application. Likewise, the author-developed AICOS and AICES should be considered preliminary instruments requiring additional psychometric evaluation in independent samples.

From a practical perspective, the findings support a balanced approach to integrating generative AI into chemistry teacher education. Rather than assuming that AI use necessarily undermines critical thinking, educators should focus on developing students' ability to use AI critically, reflectively, and appropriately within disciplinary learning contexts. Future research should further validate the measurement instruments, employ longitudinal or experimental designs, and investigate how different patterns of AI use influence higher-order thinking in chemistry education.

Acknowledgements: The authors gratefully thank to all individual participants used in the study to obtain research results

Author Contributions: Conceptualization, Kamil R. Kabzhalelov, Zhanar K. Korganbayeva; Formal analysis, Marzhan A. Nurbekova; Investigation, Kamil R. Kabzhalelov; Methodology and Supervision, Kamil R. Kabzhalelov; Writing – original draft, Marzhan A. Nurbekova; Writing – review & editing, Kamil R. Kabzhalelov, Zhanar K. Korganbayeva.

Funding: Not applicable.

Conflict of interest: The authors declare have no conflict of interest.

Disclosure of the use of Generative AI assistance:

The authors used ChatGPT (OpenAI, <https://chatgpt.com>) and Claude (Anthropic, <https://claude.ai/>) during the preparation of this manuscript to assist with English language editing,

improvement of academic writing style, literature search support, literature analysis, and the organization of scientific content. AI-generated outputs were used solely as drafting and editorial support and were critically reviewed, evaluated, and revised by the authors. All references suggested or discussed with the assistance of AI were verified against the original publications before inclusion in the manuscript. The authors take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Data Availability Statement: The data presented in this study may be obtained on request from the corresponding author.

REFERENCES

1. Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Susan Fauer Company.
2. Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 43. <https://doi.org/10.1186/s41239-023-00411-8>
3. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.
4. Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). American Philosophical Association.
5. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6. <https://doi.org/10.3390/soc15010006>
6. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage.
7. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
8. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
9. Kline, R. B. (2015). Principles and practice of structural equation modeling (4th ed.). Guilford Press.
10. Kobylarek, A., Blaszczyński, K., Słosarz, L., & Madej, M. (2022). Critical Thinking Questionnaire (CThQ): Construction and application of critical thinking test tool. *Andragogy Adult Education and Social Marketing*, 2022(1). <https://doi.org/10.15503/andr2022.1>
11. MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130–149. <https://doi.org/10.1037/1082-989X.1.2.130>
12. Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. SSRN Working Paper. <https://doi.org/10.2139/ssrn.4475995>
13. Murphy, D. H. (2023). Strategic offloading: How the value of to-be-remembered information influences offloading decision-making. *Applied Cognitive Psychology*, 37(4), 749–767. <https://doi.org/10.1002/acp.4057>
14. Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.
15. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
16. Strzelecki, A. (2024). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, 32(9), 5142–5155. <https://doi.org/10.1080/10494820.2023.2209881>

17. Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
18. Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
19. Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
20. Tian, J., & Zhang, R. (2025). Learners' AI dependence and critical thinking: The psychological mechanism of fatigue and the social buffering role of AI literacy. *Acta Psychologica*, 260, 105725. <https://doi.org/10.1016/j.actpsy.2025.105725>
21. Tiruneh, D. T., De Cock, M., & Elen, J. (2018). Designing learning environments for critical thinking: Examining effective instructional approaches. *International Journal of Science and Mathematics Education*, 16(6), 1065–1089. <https://doi.org/10.1007/s10763-017-9829-z>
22. Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.
23. Yavich, R. (2025). Will the use of AI undermine students' independent thinking? *Education Sciences*, 15(6), 669. <https://doi.org/10.3390/educsci15060669>
24. Zakaria, N. K., Hashim, H., & Jamaludin, K. A. (2025). Exploring the impact of AI on critical thinking development in ESL: A systematic literature review. *Arab World English Journal*.
25. Yilmaz, R., & Yilmaz, F. G. K. (2023). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. *Computers in Human Behavior: Artificial Humans*, 1(2), 100005. <https://doi.org/10.1016/j.chbah.2023.100005>