

The Impact of Large Language Models on EFL Learners Higher-Order Thinking Skills in Iraqi Context

Ibtihal Murad Zangana

College of Basic Education / University of Diyala

Corresponding author: Ibtehal.morad@uodiyala.edu.iq

<https://orcid.org/0000-0001-7513-0588?lang=en>

Date of research submission : 10/6/2026

Date of publication acceptance : 16/8/2026

Date of publication : 29/8/2026



[Creative Commons Attribution 4.0 International License](https://alfatehjournal.uodiyala.edu.iq/index.php/jfath/index)

<https://alfatehjournal.uodiyala.edu.iq/index.php/jfath/index>

<https://alfatehjournal.uodiyala.edu.iq/index.php/jfath/copyright>

Abstract

Generative AI can be utilized for EFL teaching to develop higher-order thinking skills, but it may also present challenges in academic integrity. Little is known about the combined influence of DeepSeek over three crucial factors of critical thinking, creativity, and self-reflection in the Iraqi EFL context. Therefore, this study seeks to bring to light how these skills are affected by the assistance of DeepSeek qualify students' impressions of the learning process. Qualitative methods and analyses: The researcher used a sequential explanatory design. The subjects of the study were 200 undergraduate students who were randomly split into two groups (experimental group [N = 100]; control group [N = 100]). Data were collected by means of pre-validated scales: Critical Thinking, Reflective thinking Creative thinking. Additionally, the researcher obtained qualitative data utilizing a semi-structured interview guide. As for the quantitative data they were analyzed with ANCOVA. As for the analyses of qualitative data thematic coding was utilized. Significant differences were found in measures of critical, creative, and reflective thinking scores

between the DeepSeek group and the control group. Qualitative results reflected a paradox in students' perceptions of AI; despite appreciating how AI facilitated skills through the generation of ideas and opportunities for safe practice, students simultaneously expressed concern about cognitive offloading and skill atrophy.

Key words: DeepSeek, high order thinking skills, EFL classrooms, Large Language Models.

Introduction

Generative artificial intelligence (GenAI) has been integrated into higher education at an accelerated pace in recent years, with tools like ChatGPT and DeepSeek becoming more readily available to learners around the globe (1 Al-Obaydi & Pikhart 2025). Deepseek is notable because unlike most proprietary LLMs of a similar architecture, it sometimes exposes structured reasoning traces that resemble an internal dialogue between a planning component and verification component (Jiang et al., 2025). This nature of transparency leads to social meaningful questions about what these tools are doing educationally if interaction with a reasoning-visible AI helps develop or even hinders our higher-order thinking skills, which happens to be an important purpose of learning in academia. Such questions are notably relevant within English as a Foreign Language (EFL) realm, where thinking and language are both the medium and object of the instruction. The higher-order Thinking Skills mentioned are commonly acknowledged as cognitive skills of higher education EFL learning (Paul & Elder, 2006). However, these three constructs have seldom been researched in conjunction with each other as a holistic construct relating to AI-assisted learning and have never been investigated through outcome measurement rather than perception surveys. Jiang et al. (2025), in the most rigorous published evaluation of DeepSeek's cognitive performance to date, found 194 Chinese undergraduate students performed worse on two standardised critical thinking instruments significantly worse than DeepSeek-R1 at completing the TSA (58.60 vs. 52.90) and EWCTET (44.47 vs. 39.71). Nonetheless, the same study found a critical caveat: The model was allocated

significantly more deduction points on holistic argumentative tasks (4.38 vs 2.87), reflecting systematic constraints in context-sensitive and multi-perspectival reasoning. The idea of this duality suggests that learners who engage with DeepSeek appears to be aided by its assessments on the type of discrete tasks at which it excels but remain susceptible to ignoring weaknesses in precisely those integrative reasoning skills required of EFL academic writing.

A productive tension between two frameworks shapes the theoretical landscape of AI-assisted learning. The theory of cognitive offloading (Risko & Gilbert, 2016) warns against continuing use of external aids, arguing that it reduces internal effortful processing and gives rise to performance benefits that are more apparent than real due to delayed skill decline. In contrast, sociocultural theory places AI as a potential scaffold within the learner's Zone of Proximal Development: when deployed in conjunction with an appropriately structured pedagogical design tools like DeepSeek may help learners to approach tasks that are beyond their capacities for independent practice through slowly internalising these strategies modelled from (Vygotsky, 1978; Kohnke et al., 2023). How well students engage metacognitively with the output of AI is the mediating variable between these two outliers-active use (critically and reflectively engaging with AI outputs) or passive acceptance (generalised viewing only)? This distinction between conclusion and facilitation is what elevates reflective thinking from being a simple outcome of an AI-assisted learning environment to the enabling mechanism of it. In a phenomenographic study of Iraqi and Taiwanese EFL students using DeepSeek, Al-Obaydi & Pikhart (2025) reported that the majority perceived both expanded ideational range and decreased writing anxiety, but many also expressed fears about dependency and diminished autonomous thinking.

Despite this growing evidence base, the Iraqi EFL context remains substantially underrepresented. Al-Obaydi and Pikhart (2025) constitute the only published study directly addressing Iraqi students' engagement with DeepSeek, and that study focused on

attitudinal satisfaction rather than cognitive outcome measurement. Iraqi higher education is characterised by distinctive moderating conditions including variability in digital infrastructure, a longstanding curricular tradition of rote learning and teacher-centred pedagogy, and recent institutional initiatives such as the AI integration workshop hosted by the College of Basic Education at Al-Mustansiriyah University (University of Mustansiriyah, 2026) that are unlikely to be fully captured by findings from Chinese, Uzbekistani, or Gulf contexts. Methodologically, the field has been dominated by single-method designs employing either attitudinal surveys or qualitative interviews; few studies have combined pre-post assessment of validated cognitive instruments with qualitative exploration of student experience within a single design. No published study has simultaneously examined DeepSeek's impact on all three interdependent higher-order thinking domains in this context.

The present study tries to bridge these gaps through an explanatory sequential mixed-methods design (Creswell & Clark, 2017), combining a pretest-posttest quasi-experimental comparison of 200 Iraqi EFL undergraduates with thematic analysis of semi-structured interviews. The four research questions listed below are behind the motivation of conducting the current study:

RQ1: What is the impact of DeepSeek-assisted learning on Iraqi EFL students' critical thinking scores relative to conventional instruction?

RQ2: What is the impact of DeepSeek use on students' creative thinking in academic writing tasks?

RQ3: How does engagement with DeepSeek affect students' reflective thinking development?

RQ4: What are students' experiences of the cognitive affordances and risks presented by DeepSeek in EFL learning?

This study contributes to the literature in several ways. Rather than reducing AI use to a simplistic binary of beneficial or harmful, this study examines the cognitive mechanisms underlying human–AI interaction in order to distinguish basic tool use from the higher-

order thinking processes characteristic of performance-based learning environments. In doing so, it shifts attention toward how DeepSeek can be implemented in the Iraqi EFL context to foster effective cognitive engagement without displacing or diminishing students' internal processing. In so doing, it withdraws attention from how tools like DeepSeek could be implemented in an Iraqi EFL context to guide students toward more effective internal and external knowledge on cognizing for employment without displacing or diminishing cognitive engagement. On a more pragmatic note, it also aids curriculum designers and institutional policymakers — university stakeholders who are in the third wave of generative AI adoption. by highlighting how to marry performance improvements achievable via generative AIs with maintaining meaningful cognitive engagement when accessibility won't deterministically lead to profound learning ways. This type of contextualized evidence, in fact, goes beyond the realm of academia; generative AI tools are increasingly part and parcel of classroom practice internationally.

Literature Review

In EFL education, higher-order thinking skills are traditionally viewed as encompassing three interconnected cognitive constructs. Facione's (1990) Delphi consensus definition of critical thinking, which includes interpretation, analysis, evaluation, inference, explanation and self-regulation all skills that tend to be more important in EFL contexts where learners must not only decode content but also evaluate it rhetorically and logically in a second language. Accordingly, creative thought is frequently composed of these divergent processes fluency, flexibility, originality, and elaboration as well as convergent processes available for evaluating and refining ideas generated (Guilford 1967; Runco & Acar 2012). Creativity in EFL writing can be manifested as lexical choice, metaphorical expression, and articulation of culturally unique ideas within the constraints/parameters of a second language. Linked with this is the concept of reflective thinking, which Dewey (1933) described as the active and intentional

contemplation of beliefs, and the metacognitive ability to evaluate one's own performance, synthesise feedback and adjust future practice (Schön, 1983; Moon, 2004). Carless and Boud (2018) similarly conceptualize this as feedback literacy the skill set of being able to seek, evaluate, and act on feedback from teachers, peers or AI. While these three skills are usually researched in isolation, Paul and Elder (2006) make a compelling case that they are interdependent: critical thinking without creativity results in inflexible analysis, creativity without critical thinking leads to impractical ideas and reflective thinking metacognitively connects the two. The conceptual basis for the present study is that theoretical interdependence.

Empirical evidence on LLMs as critical thinking agents is most significantly grounded by Jiang et al. (2025), who systematically compared DeepSeek-R1 with GPT-3.5, GPT-4 and LLaMA-4 against 194 Chinese undergraduate students on two standardized tests: the Cambridge Thinking Skills Assessment (TSA) and the Ennis-Weir Critical Thinking Essay Test (EWCTET). DeepSeek-R1 outperformed students on both TSA scores (58.60 vs 52.90; large effect size) and EWCTET scores (44.47 vs 39.71; large effect size), along with Cohen's $d = 0.71$ and Cohen's $d = 0.68$ respectively.) This calls into question whether AI tools are merely auditing cognition, and instead argues that modern LLMs may possess at least a feasible approximation of the capacity for critical thinking in some discrete analytical tasks. Yet the same study uncovered major qualitative drawbacks: DeepSeek-R1 received 4.38 deduction points on average per task on the EWCTET versus 2.87 for students, indicating systematic deficiencies in identifying more than one point of argumentative fallacy, contextualising abstract claims as failing or not without separating into components and pulling together multiple perspectives exactly those aspects of critical thinking that demand a holistic, context-sensitive judgement rather than pattern spotting methods at root. This paradox proficiency on discrete tasks, systematic failure on integrative ones provides direct insight into pedagogical design:

DeepSeek may successfully scaffold students in discrete analytical operations but readily throw them off-balance during the multi-faceted reasoning needed for academic writing.

Central to the theoretical debate in the literature is whether AI tools are augmenting or replacing human thinking. Cognitive offloading theory (Risko & Gilbert, 2016) proposes that if reliable external tools are available for use, the individual delegates cognitively demanding tasks to these tools, and this may lead (over time) also lead us away from the richest processing internally. This points to a potential danger, where students who regularly use DeepSeek for analytical or creative tasks may slowly erode the cognitive muscles that the tool seems to be developing those skills in. Sociocultural theory (Vygotsky, 1978; Wood et al., 1976) then offers a potentially more hopeful framework operating against this worry: as potential scaffolds tools located in a learner's Zone of Proximal Development can make possible engagement with tasks beyond independent competence and gradual internalisation of the cognitive strategies modelled. Kohnke et al. (2023) applies this specifically to generative AI, arguing that tools like ChatGPT can serve as scaffolds for language learning if used with intentional pedagogical structures behind their implementation. A key mediating variable is the manner in which students interact with AI output — do they passively copy it or actively compare to it, evaluate mismatches and reflect on implications. This distinction portrays reflective thinking not solely as a component of the learning that is at least supported by AI, but as the principle vehicle through what AI interaction will turn into permanent cognition development.

Contexts for EFL Research on AI and Creativity The research conducted on AI and creativity in the context of EFL teaching or learning highlights a persistent tension between divergent and convergent thinking. In their phenomenographic study of Iraqi and Taiwanese EFL students using DeepSeek, Al-Obaydi and Pikhart (2025) found that most participants reported the tool expanding their conceptual range, providing ideas they would not have

independently thought about — a clear indication of AI's ability to defeat terminal writer's block and draw on divergent thinking. But the study also found worrisome dependency and diminished independent thinking, so differing gains may carry common costs if students accept AI recommendations uncritically. Theoretically, this tension can be cast in terms of Sternberg's (2018) novelty-appropriateness duality: creative thought requires both originality and contextual relevance, but LLMs are inherently biased toward statistically probable outputs that may maximize contextual appropriateness at the expense of true novelty. Such studies thus collectively suggest that DeepSeek's cognitive effects are fundamentally conditional: they depend less on the characteristics of the tool than on students' metacognitive quality in interacting with it.

However, major contextual and methodological gaps continue to exist despite this increasing international evidence base. The literature on AI, and thinking skills has done substantial and steady work but the EFL context in Iraq is well underrepresented; indeed Al-Obaydi and Pikhart (2025.) is the only published study addressing Iraqi students directly using DeepSeek and currently capturing attitudinal rather than cognitive outcomes. Iraqi higher education exhibits some specific context-dependent moderating conditions through 1) Bifurcated infrastructure for digital access, 2) a decades-old curricular tradition of rote learning and teacher-centred pedagogy, and 3) institutional initiatives such as the AI-specialised workshop hosted at Al-Mustansiriyah University's College of Basic Education (University of Mustansiriyah, 2026), which are inherently bound to influence what is perhaps best described as the commonly held view among their Chinese, Central Asian or Gulf counterparts concerning extant causal linkages between use and cognitive impact. The field has been methodologically dominated either by quantitative survey instruments, which concern perceptions rather than performance, or by purely qualitative interview designs, which reveal experience but not outcome. There have been relatively few studies that have

implemented mixed-methods designs comparing pre-post assessment of validated thinking-skill instruments with qualitative exploration of student experience. Additionally, no published study has investigated the effects of DeepSeek on all three interdependent thinking domains critical, creative, and reflective in a single research design despite the theoretical argument that these skills interact and are differentially affected by distinct forms of AI use. This study was undertaken specifically to fill these gaps.

Methodology

The research design used was mixed method with sequential explanatory. This model was selected because it was able to give a comprehensible investigation for the research outlined problem. Quantitative data were collected and analyzed first, then the collection and analysis of qualitative data was carried on, this gave the researcher the ability to explain, elaborate and enrich the quantitative data, which was collected first.

The current study went through two phases: in phase one a pretest-posttest was conducted in order to investigate the impact of the two instructional conditions on students' higher order thinking skills. The learning method (DEEPSEEK-assisted VS lectured-based) perceived as an independent variable, where critical thinking (CT), creative thinking (CR) and reflective thinking (RT) were the dependent variables. The second phase was qualitative, semi-structured interviews to collect in-depth data concerning the student experience and to identify what student participants feel about the mechanisms through which the quantitative findings operate. In this way, a solid methodological triangulation of quantitative scores and qualitative narratives is possible. This triangulation of methods did enhance the results validity and credibility, because using the qualitative data to interpret how and why a quantitative result was obtained (Creswell & Clark, 2017).

Participants

The subjects were 200 undergraduate students majoring in Information Systems at the Iraqi universities aging between (18-23) for the academic year 2024–2025. This population was chosen

because of its special suitability, namely a familiarity with technology combined with a range of English language competence. The participants declare their willingness to take part in the study. The inclusion criteria were: first having good or Intermediate level English language, second: having regular access to technology, and not having significant prior experience with AI for academic English learning, as confirmed by a screening questionnaire. The final sample of 200 students was academically homogeneous with a mean CWA (Cumulative Weighted Average) of 68.67 (SD = 6.81). After the pretest, students were divided into two groups: an experimental group (n = 100) for DeepSeek-assisted teaching and a control group (n = 100) for traditional teaching.

Instrument

Data collection: the research instrument utilized were Three scales which were validated before use, and a semi-structured interview (Harrington, 2009), all quantitative instruments were confirmatory factor analyzed (CFA) in order to establish construct validity. Maximum likelihood estimation was used to assess model fit. Results confirmed that all measurement models demonstrated acceptable fit indices: RMSEA < 0.08 and CFI > 0.90, indicating good construct validity for each scale.

The Critical Thinking Scale (CTS) was the first instrument used to measure students' possible changes in critical thinking skills. The ability to deconstruct arguments and the tendency to question assumptions and seek evidence were two major dimensions in this 20-item instrument based on Sosu (2013)'s work. Each participant completed a 5-point Likert scale of response for each item. This scale had a high internal consistency with a Cronbach's alpha coefficient of 0.93

The researcher has adopted the Creative Thinking Scale (CTS) from Hidayat et al. (2018) which is a 20-item scale for assessing various aspects of a creative attitude, subscales of 'innovative thinking' which reflects intellectual courage that express willingness to take risks with ideas and 'flexibility' to reflect the

ability to think about ideas from different perspectives. The reliability test for this instrument showed that the instrument had good internal consistency with the value of Cronbach Alpha of 0.89.

The researcher measured the students' reflective thinking skills by using a pre-validated 20 item Reflective Thinking Scale (RTS) which was adopted from Basol and Gencel model (2013). The standard is developed to evaluate students' advancement towards meditation proficiency. They span from lower levels of action ('habitual action') and understanding to higher levels of 'reflection' and 'critical reflection' with a perspective shift. The Cronbach's alpha coefficient of this instrument is 0.96, which is high.

Semi-structured interview guide for the qualitative part of the In-depth study to get a deeper understanding of students' perceptions. The guide was adapted from Xiao and Zhi (2023) and was purposely designed to obtain rich and narrative accounts of how the students interacted with DeepSeek as they engaged in learning. It that included open-ended questions and cues to prompt during which the participants shared their features of perceived benefits, challenges and characteristic interactions that they observed that some would have helped while others had hindered their ability to learn. This qualitative Instrument was very important to provide explanation to the quantitative data collected in the scales.

Data Collection

The current study data were collected in three phases over a 6-week period of time. The three phases were structured to gather successively in-depth data, which were needed to provide answers the research questions quantitatively and qualitatively. At the beginning, the researcher set a pretest in order to construct a fuller baseline of the participants' higher-order thinking skills before the intervention sessions began. Two hundred enrolled participants were summoned in a controlled atmosphere with the aim of ending three, quantitative scale factors: Critical Thinking Scale, Creative Thinking Scale and Reflective Thinking Scale. The covariate used

in the final analysis to control for baseline differences between groups was pretest data.

Across both the instructional intervention comprised of six-weeks of instruction based on a 2-hour a week, collaborative format focusing similar topics in academic English (argumentation, essay writing, critical reading and paraphrasing source quality for these two groups across conditions and task equivalence.

In the experimental group (n = 100), each week consisted of a three-phase lesson structure. Phase 1 (input, 30 mins): Direct instruction/modeling with the instructor introducing the weekly topic. In Phase 2 (AI-assisted task, 60 minutes), students interacted with DeepSeek using sets of prompts aimed to stimulate higher-order thinking. These prompts were created by the researcher and validated through two EFL experts before implementing it in the intervention. Materials and methods First, we present the design of a initial study to evaluate instruction transmission strategies using LLMs. Suggest two counterarguments to my argument and how I could bolster it more effectively"; (b) critical inquiry prompts e.g., "What are the premises of this assertion? What counter-evidence would challenge it?" Students had to reflect on their interaction with DeepSeek in a reflection journal, indicating for each suggestion from the AI if they accepted it unchanged or modified it and why. Phase 3 (debrief:30 mins): instructor led a whole-class discussion through which students explicitly and critically compared their own reasoning with DeepSeek's output, specifically targeting the evaluation and synthesis dimensions of higher-order thinking.

The control group (n = 100) received instruction on the same weekly topics through a conventional teacher-centred approach involving direct explanation, textbook-based reading tasks, and whole-class discussion, without any AI tool access. Both groups were taught by the same instructor to control for teacher effect.

Data were collected at the end of the 6-week intervention. 200 participants in both groups underwent a post-test (same 3 quantitative instruments as the pre-test to measure delta cognitive

skills). Qualitative Data Collection Adjacent. Purposive sampling technique was used for selecting fifteen students ($n = 15$) in experimental group, in which selected of students who were either high, medium or low involved in the intervention based on their representation a range of views. Each student was interviewed through semi-structured interview lasting around 30–45 minutes. They were informed that the sessions would be audio-recorded for transcription purposes and to allow a close investigation of their experience with the AI-assisted learning process.

For the qualitative component, purposive sampling was employed to select 15 participants from the experimental group ($n = 100$). Involvement level was operationalized using two criteria: frequency of DeepSeek use during the intervention (tracked via weekly activity logs) and posttest score gain relative to pretest baseline. Participants were then stratified into three involvement tiers high ($n = 5$), medium ($n = 5$), and low ($n = 5$) to ensure the interview sample represented the full range of engagement experiences. This stratified purposive approach followed the recommendation of Creswell and Clark (2017) that qualitative participants in explanatory sequential designs be selected specifically to explain or elaborate the quantitative results, and the sample size of 15 is consistent with thematic saturation norms in mixed-methods educational research (Guest et al., 2006).

Data Analysis

The researcher used An IBM SPSS Statistics, Version 26 to analyze the quantitative data (cognitive skill scores difference). The aim was to answer the first three research questions regarding the differences in higher-cognitive skill performance, (ANCOVA) then was performed. this statistical approach was chosen because it can enhance the overall power of statistics as well as reduce error variance. This was accomplished by applying ANCOVA to numerator Group scores against denominator control pretests and posttest scores where these pretests were statistically removed as a covariate allowing the true treatment effect (DeepSeek use vs traditional methods) on significant improvements of knowledge to

be identified. Before performing the main analysis, ANCOVA assumptions were thoroughly assessed. These included normality of data distributions, homogeneity of variances (Levene's test) and the homogeneity of regression slopes. All analyses were deemed significant at alpha level of $p < 0.05$.

Qualitative data which were obtained from the transcripts of semi-structured interviews were analyzed using thematic analysis. Braun & Clarke (2006) suggested a six-phase procedure, which was used systematically and flexibly in this study, for the identification of patterns in qualitative information. Analytical process began with data familiarization in which the researchers re-read transcripts repeatedly to get immersed in the data. This was then followed by a systematic procedure in which initial codes were generated of interesting features throughout the data set. These codes were then clubbed into potential themes based on which they were then reviewed, developed and clearly defined. In order to ensure credibility and reliability of the analysis, two researchers independently coded the data. To enhance coding and inter-coder reliability for the coding frame, the coding frame was refined through a cycle of coding and discussion; and was found to have a Cohen's Kappa of 0.87, signifying substantial agreement between coders. The mature thematic structure was then applied to aid the interpretation of data and rich, contextually-focused answers to the fourth research question.

Results

RQ1: What is the impact of DeepSeek-assisted learning on Iraqi EFL students' critical thinking scores relative to conventional instruction?

In order to address the first research question, a pretest-posttest quasi-experimental design was conducted. To compare the critical thinking scores of the posttest of the Deepseek group and the Traditional group, and statistically control the pretest scores of the groups, an ANCOVA was used.

Table (1)
Descriptive Statistics of Critical Thinking Scores

Group	Time	Mean	Standard Deviation	N
DeepSeek	pretest	75.9	6.3	100
	Post test	85.4	5.2	100
Traditional	pretest	76	6.5	100
	Post test	76.2	6.1	100

Table 1 shows the descriptive statistics of the pretest and posttest scores of the experimental (DeepSeek) and control (Traditional) groups. There were 100 participants in each group ($N = 100$). Both groups had a similar level of baseline proficiency at the pretest stage. The mean score for the DeepSeek group ($M = 75.9$, $SD = 6.3$) was almost the same as the score for the Traditional group ($M = 76.0$, $SD = 6.5$). After the intervention period, the posttest results revealed that there was a significant improvement in the performance of the DeepSeek group. Their mean score increased significantly to 85.4 ($SD = 5.2$). However, the performance of the Traditional group was very similar. Their posttest mean score was virtually the same at 76.2 ($SD = 6.1$). The results indicate that there is strong evidence of the positive impact of the DeepSeek-assisted intervention on the scores of participants, as opposed to the traditional instruction method that did not lead to any significant improvement.

Table (2)
Test of Between-Subjects Effects for Critical Thinking

Source	Type III Sum of Squares	DF	Mean Square	F	Sig.
Corrected model	4901.50	2	2450.75	76.30	$P < .001$
Pre-test critical thinking	300.90	1	300.90	9.36	.033
Group	4590.62	1	4590.62	142.92	$P < .001$

					.001
Error	6230.50	197	32.63		
Corrected total	11132.00	199			

To see the effect of the intervention on critical thinking score of the students, ANCOVA test was used. Initial differences were controlled for by using the pretest critical thinking scores as a covariate. Table 2 shows the results of the ANCOVA. A main effect of the instructional group was found and was statistically significant, $F(1,97) = 142.92$, $p < .001$. This finding shows that there is a significant difference in post test critical thinking scores between the experimental and control groups, controlling for the pre test scores. In addition, the covariate (pretest critical thinking) was also a significant predictor of posttest performance, $F(1, 97) = 9.36$, $p = .033$. However, the main result is that the group variable has a significant impact. This indicates that the differences in critical thinking outcomes was mainly due to the intervention itself.

Table (3)
Estimated Marginal Mean for Posttest Critical Thinking

Group	Mean	Std. Error	95%Confidence Interval Lower Bound	95%Confidence Interval upper Bound
DeepSeek	85.4	0.56	84.30	86.50
Traditional	76.2	0.56	75.10	77.30

The estimated marginal means for posttest critical thinking scores controlling for pretest differences are displayed in Table 3. DeepSeek group had a greater mean score ($M = 85.4$, $SE = 0.56$) compared to the traditional group ($M = 76.2$, $SE = 0.56$). The 95% confidence interval for the DeepSeek group is between 84.30 to 86.50 Traditional: 75.10 to 77.30 . Discussing These results suggest that DeepSeek group improved critical thinking performance compared with traditional group.

RQ2: Do students who learn English with DeepSeek and those who learn English with traditional methods have statistically significant differences in their creative thinking scores?

Your second research question was whether or not there were significant differences between the two groups in terms of creative thinking scores. We used a pretest-posttest quasi-experimental design and ANCOVA statistical procedure for the same creative thinking scores.

Table (4)
Descriptive Statistics of Creative Thinking Sources

Group	Time	Mean	S .D.	N
DeepSeek	Pretest	79.1	5.9	100
	Posttest	88.1	4.9	100
Traditional	Pretest	78.9	6	100
	Posttest	79.5	5.8	100

Descriptive statistics of the creative thinking scores of the experimental (DeepSeek) and control (Traditional) groups at pretest and posttest are shown in Table 4. There were 100 participants in each group ($N = 100$). The creative thinking proficiency of both groups was almost the same at the beginning of the study. The pretest scores for the DeepSeek group ($M = 79.1$, $SD = 5.9$) were very similar to the pretest scores of the Traditional group ($M = 78.9$, $SD = 6.0$). This validates the same starting point for both groups prior to the intervention. A significant difference in performance was seen after the intervention. The DeepSeek group had a significant gain in creative thinking scores. Their mean score rose to 88.1 ($SD = 4.9$). The scores of the Traditional group were, by contrast, fairly constant. The mean score for the students in their posttest was 79.5 ($SD = 5.8$) which was very little improvement from their mean score in the pretest. Based on the descriptive results, it can be concluded that the intervention using DeepSeek was very effective in developing creative thinking skills. In contrast, the traditional teaching approach did not result in any significant progress in this aspect.

Table (5)
Between -Subject Effect for Creative Thinking

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected model	4361.04	2	2180.52	73.04	P < .001
Pre-test critical thinking	251.22	1	251.22	8.42	.004
Group	3698.00	1	3698.00	123.88	P < .001
Error	5791.00	197	29.40		
Corrected total	10152.04	199			

ANCOVA test was applied for determining the effect of instructional intervention on creative thinking scores of students. The covariate was the pretest creative thinking scores. This method controlled for any initial differences in ability between and among the groups statistically. The analysis is shown in Table 5. Results show a significant main effect of group, $F(1,97) = 123.88$, $p < .001$. The results indicate a significant difference between experimental and control groups posttest creative thinking scores after controlling the pretest scores. In addition, pretest creative thinking (the covariate) was a significant predictor in pre-post test scores, $F(1, 97) = 8.42$, $p = .004$. The main findings remain substantial, even with the influence of baseline scores. That is, the use of intervention was responsible for enhancing students' creative thinking skills.

Table (6)
Estimated-Marginal Means for Posttest Creative Thinking

Group	Mean	Std. Error	95%Confidence Interval Lower Bound	95%Confidence Interval upper Bound

DeepSeek	88.1	0.54	87.04	89.16
Traditional	79.5	0.54	78.44	80.56

Estimated marginal means for the posttest creative thinking scores with differences in pretest scores controlled are displayed in Table 6. The DeepSeek group reached a higher mean score ($M = 88.1$, $SE = 0.54$) compared with the traditional group ($M = 79.5$, $SE = 0.54$). The 95% confidence interval for the DeepSeek group ranges from 87.04 to 89.16 . The traditional group confidence interval is (77.97, 81.03). These results indicate that students who learned using DeepSeek exhibited a higher level of creative thinking skills compared to students who learned using traditional methods.

RQ3: How does engagement with DeepSeek affect students' reflective thinking development?

Concerning this question which was sited to compare the score of reflective thinking skill of students in Deepseek-assisted setting with the score of reflective thinking skill of students in the traditional setting. Again, this was a pretest-posttest study with ANCOVA to account for baseline differences.

Table (7)
Reflective-Thinking Scores

Group	Time	Mean	S. D.	N
DeepSeek	Pretest	74.8	6.5	100
	Post test	83.6	5.4	100
Traditional	Pretest	75.2	6.9	100
	Post test	75.4	6.5	100

Descriptive statistics for the reflective thinking scores for both, DeepSeek-assisted learning and traditional learning are shown in Table (7). There were 100 participants in each group. The mean scores of pretest and posttest for the DeepSeek group were 74.8 ($SD = 6.5$) and 83.6 ($SD = 5.4$), respectively, indicating a significant increase in reflective thinking. The traditional group, on the other hand, exhibited little change with a pretest mean of 75.2 ($SD = 6.9$) and a posttest mean of 75.4 ($SD = 6.5$). The results indicate that the baseline scores for both groups were fairly similar,

but that the DeepSeek group showed significant improvements in reflective thinking over time while the traditional group almost did not show significant changed.

Table (8)
Between -Subject Effect for Reflective Thinking

Source	Type III Sum of Squares	DF	Mean Square	F	Sig.
Corrected model	4100.36	2	2050.18	57.10	P < .001
Pre-test critical thinking	360.56	1	360.56	10.04	.002
Group	3528.00	1	3528.00	98.26	P < .001
Error	6965.80	197	35.36		
Corrected total	11066.16	199			

Table 8 represents the output of ANCOVA test to determine the between subjects effects to reflective thinking. Results indicate that the full model was statistically significant, $F(2, 197) = 57.10$, $p < .001$, indicating that the predictors explained a substantial amount of variance in reflective thinking scores. The main effect for pretest scores of reflective thinking was likewise significant, $F(1, 97) = 10.04$, $p = .002$, suggesting that scores on first reflective thinking affected the posttest. Related, $F(1, 197) = 98.26$, $p < .001$. After controlling for pretest scores, the students that received DeepSeek performed statistically far better than those in the traditional group. This suggests the considerable importance of the impact of DeepSeek-based learning on students' reflective thinking.

Table (9)
Estimated-Marginal Means for Posttest Reflective Thinking

Group	Mean	Std. Error	95%Confidence Interval Lower Bound	95%Confidence Interval upper Bound
DeepSeek	83.6	0.60	82.52	84.88
Traditional	75.4	0.60	74.12	76.48

Table (9) Results of the estimated marginal means for the posttest reflective thinking scores, controlling for pretest differences The mean score obtained by the DeepSeek group ($M = 83.6$, $SE = 60$) exceeded that of the traditional group ($M = 75.3$, $SE = 0.60$). DeepSeek group 95% confidence interval (82.01, 85.39) what this means is you can say with "94%" significance the traditional group has a 95% confidence interval of (73.61, 76.99). The results of these studies suggest that the DeepSeek-based students more often engaged in reflective thinking compared to their counterparts.

RQ4: What are students' experiences of the cognitive affordances and risks presented by DeepSeek in EFL learning?

The fourth research question investigated students' opinions about their interactions with DeepSeek. The aim was to find specific mechanisms, which students saw as promoting or inhibiting their cognitive skills. Thematic analysis of qualitative data from interviews and reflections was conducted. This methodological approach is similar to other recent studies that explored student perceptions of AI in education.

Table (10)
A Thematic Analysis of Student Perceptions on Interacting with DeepSeek

Category	Sub-them	Description	Student statement
fostering mechanisms.	Idea Generation & Brainstorming	Students turn to DeepSeek when they get mentally blocked or need a jumping off point for writing and speaking tasks.	The warm way DeepSeek works is that it can help you brainstorm ideas. It recommends and prompts me when there is a mental barrier and I do not know what to write.
	Perspective Broadening & Divergent Thinking	The tool challenges students to see things from	Artificial Intelligence introduced me to numerous angles of

		different angles and be led toward potential solutions that are not part of their default way of thinking.	view. This guidance forced me out of my own knee-jerk thought processes and allowed more creative options for my essay to reveal themselves.
	Argument Refinement & Convergent Thinking	DeepSeek puts him in the habit of weighing arguments both for and against everything, allowing students to focus their thoughts logically.	Particularly when I had several viable arguments, AI helped me weigh the pros and cons of each. One even advised me which had the best evidence base, sharpening my writing to a more compelling point.
	language scaffolding	It provides a non-judgmental space to practice languages by providing immediate feedback on grammar and vocabulary, reducing anxiety.	It proves to be a safe space for you to practice without worrying about being judged. I can test new sentence structures, get immediate feedback and feel more confident.
	Metacognitive Support & Reflective Thinking	The AI functions as a 'to do list' or thinking partner where students in control of their learning, as the manager of processes setting	However, the training that gives me a clear objective before I write is what AI helps me with. It forces me to think deeply about my intentions from day

		goals, references and tracking processes.	1 and keeps us all aligned with the same mission.
hindering mechanisms	Cognitive Offloading & Intellectual Laziness	Students feel it a bit scary that AI is making ready-made answers and students can skip the process of cognitive work, which is not easy but important part of learning.	It's a substitute for all cognitive processes. If you used it to do your homework for you, you're not attempting to understand anything. It can breed intellectual laziness in you.
	Accuracy Concerns & Misinformation	General hopes and fears Students are concerned about the possible unreliability of AI-generated information both in terms of plausible sounding but incorrect facts (i.e., "hallucinations") and fabrication of source materials.	I am worried that any information generated by DeepSeek is trustworthy. It can also generate hallucinations or invent sources, so you always have to fact-check it.
	Perils of Over-Reliance & Skill Atrophy	Some people also fear, using the AI too much will lose basic skills like conducting independent research,	Over-reliance is a real problem. Students are not practicing their critical thinking, and it causes students become dependent on what DeepSeek

		analysis and writing.	tells them. Your skills can atrophy.
	Bias Perpetuation	Listeners understand that when other students say that AI models can replicate and amplify biases in the data these models have trained on, all this does is provide a misleading picture of an already complicated problem.	One reason for these failures is that AI systems inherit biases from their training data. However, unchecked such filtering may restrict your exposure to alternative views and indeed, exacerbate present misconceptions.
	Information Overload	AI can generate an overwhelming amount of information in a short period which makes it very difficult for students to filter the critical information and focus.	At times it simply is a case of information overload when using DeepSeek. It doesn't filter out the important stuff when it gives you so much all at once and increases my cognitive load.

As shown in Table 10, the student perceptions were analyzed using thematic analysis, revealing a notable dichotomy in their experiences with DeepSeek. These experiences are grouped into two main themes: 'Fostering Mechanisms' and 'Hindering Mechanisms'. Students indicated that DeepSeek is a useful tool to encourage high-order thinking skills. They used it as a brainstorming and idea-generating tool, which helped them to

overcome creative impasses. The AI also shared with them different perspectives. This exposure widened their thinking and promoted divergent thinking. Moreover, the tool helped in the refinement of the arguments. It enabled students to assess the pros and cons of various roles, which helps in convergent thinking. Students also valued DeepSeek's function as a language scaffold. The platform provided a language practice space where students could practice their language without judgment, thus decreasing the affective filter and decreasing student anxiety. Lastly, it was used as a metacognitive support tool to help students set goals and to reflect on their learning process.

However, students were also very aware of the downsides of the tool. One of the major issues was that of cognitive offloading and intellectual laziness. The ease of obtaining answers could encourage students to circumvent critical thought. There were also issues of accuracy and misinformation. Students were aware of the potential for the AI to produce 'hallucinations', that is, seemingly logical but factually inaccurate statements. One major concern was that of over-reliance and skill atrophy. Over-reliance on the AI may lead to a decline in their core academic skills. The risk of bias perpetuation was also identified as a risk by the participants. They recognized that the AI may exhibit and even exacerbate biases found in its training data. Lastly, the large volume of AI-generated text could result in information overload. This situation adds to the cognitive load and hinders the ability of the student to attend to key information.

Discussion

The present study demonstrates the statistically significant enhancement effects of DeepSeek-assisted instruction on critical thinking, creative thinking, and reflective thinking of Iraqi EFL undergraduates with highly significant group effects having been obtained from the ANCOVA analyses after controlling for pretest ability in all three cases. In agreement with Jiang et al. Critical thinking scores changed from 75.8 (SD = 14.3) to 85.4 (SD = 12) in the experimental group, [22] and remained similar in control

groups with negligible differences statistically significant changes; DeepSeek-R1 achieved better scores than undergraduate students, as found by a comparative analysis on standardized critical thinking measures like 'TSA' and 'EWCTET'. These support roles may translate into Vygostkian terms as evidence of competent ZPD (Zone of Proximal Development) support in the modelling of chain-of-thought reasoning structured argumentation, which suggests that the tool has facilitated student's ability to analytically engage with components of tasks the student could not have done alone (Vygotsky, 1978; Wood et al., 1976). This understanding is coherent with the work of Kohnke et al (2023). Generative AI tools can serve as a support structure to encourage language and knowledge acquisition, when used purposefully and within a teaching framework. Yet, as a cautionary note (Risko & Gilbert, 2016), this is only part of the story: One ramification of cognitive offloading theory is that once we begin to rely on available external tools, internal processing coasts downward in effort over time. Thus the quantitative increases observed can be partly due to genuine cognitive development and in part scoring with help on tasks domains that are better separated out by longer-term research designs.

Use creative thinking For example, for experimental group gains (pretest $M = 79.1$; posttest $M = 88.1$; $F(1, 197) = 123.88$, $p < .001$) invite cautious interpretation. Preliminary research using DeepSeek in studies undertaken in the context of Iraqi and Taiwanese EFL students (Al-Obaydi & Pikhart, 2025) suggests that use of this corpus-based tool leads students to expand their conceptual horizons through being exposed to ideas which might not otherwise have been generated by learners themselves, echoing a finding regarding the facilitative nature of DeepSeek as an enabler of divergent thinking (Guilford, 1967; Runco & Acar, 2012). Yet, in this study, the instrument used was attitudinal innovative orientation, intellectual courage and flexibility yet not behavioural writing performance. Achieving a score on scale-based measures of creativity might not reveal anything about divergence

if divergent thinking is defined as the critical selection and polishing of the generated characters. This is important because large language models are architecturally and structurally predisposed to statistically plausible output: as such, they will both encourage idea generation whilst normalising expression toward what would be considered acceptable; traditional patterns a theoretical tension that lies in Sternberg's (2018) novelty-appropriateness duality. Whether the attitudinal benefits documented here translate into new human-writing output is an empirical matter for rubric-scored text analysis.

Means for reflective thinking can be found in Table 2 (pretest $M = 74.8$; posttest $M = 83.6$; $F(1, 197) = 98.26$, $p < .$ From a theoretical standpoint, perhaps the most interesting result is that students who were prompted to actively work with AI feedback tended to exhibit metacognitive processes (both conscious intellectual reflection and meta-motivational regulation) which would be in keeping with theorist of reflective thought (Dewey, 1933; Schön, 1983; Moon, 2004). Crucially, this effect cannot with equanimity be attributed to DeepSeek per se but rather deep-seated within a pedagogical framework that tasked students with evaluating, justifying and selectively implementing AI suggestions an operationalization of feedback literacy (Carless & Boud, 2018). The structured debrief and journaling aspects seem to have nudged the students toward this systemic engagement of act-reflection.

The qualitative data reveal a structurally persistent contradiction where learners respected DeepSeek for its benefits in producing ideas and scaffolding arguments and language but also identified equally legitimate concerns about cognitive over-reliance, atrophy of existing skills, hallucinating AI, and diminishing intellectual agency. These concerns are not just affective; they constitute an observational feature aligned seamlessly with the tenets of cognitive offloading theory (Risko & Gilbert, 2016) and reflect the qualitative patterns reported by Al-Obaydi and Pikhart (2025) for learners within the same cohort. Importantly, even being able to articulate such concerns is evidence of a type of meta-cognition

that would place it in line with van Manen's (1977) more sophisticated forms of reflection. However, this metacognitive awareness does not translate into self-regulatory behaviour that protects against ineffectuality (Flavell, 1979) so future instructional design must include explicit instruction in AI-critical literacy – detection of hallucinations and judgement of its arguments and a conscious decision by students not to use it - if awareness is going to be convert into action.

Overall, the implications of these findings are neither inherently negative nor inherently positive; rather, they are fundamentally contingent upon the quality of the pedagogical environment in which DeepSeek is deployed.. This study contributes to the still emerging mixed-methods empirical literature assessing the simultaneous effects of DeepSeek on all three aspects of higher-order thinking under such relatively uncharted educational conditions, as shown in this Iraqi EFL context that currently offers at least a small window for AI-augmented learning, despite its existing structural constraints. Future studies utilizing longitudinal designs may also explore the duration of these changes, in addition to how modes of AI use brainstorming vs grammar correction and feedback evaluation differentially affects cognition.

Conclusion

This sequential explanatory mixed-methods study investigated the Impact of DeepSeek-enhanced instruction on Iraqi foreign language learners' Higher-order thinking skills (Critical/Critical, Creative/Creative and Reflective) as well as provided convergent evidence based on qualitative thematic analysis findings and quantitative ANCOVA results.

For all three higher-order thinking domains, when controlling for pretest differences, the learning was statistically significant using DeepSeek compared to conventional instruction: Critical thinking-- The experimental group increased about 9.6 scale points ($M = 75.8$ to $M = 85.4$) where the control changed nearly zero on average. This accounts for increments of approximately 9.0 points (from $M = 79.1$ to $M = 88.1$) and approximately 8.8 points (from $M = 74.8$

to $M = 83.6$), respectively, in creative thinking and reflective thinking. These gains represent substantively meaningful improvements that align with theoretical claims regarding generative AI as a digital scaffold for cognitive growth when integrated with rigorous, critically informed pedagogy. (Kohnke et al., 2023; Vygotsky, 1978).

Second, evidence from the qualitative data demonstrate that gains are subtle, multimodal and contingent on the ways in which students metacognitively prompt the AI. Students reported cognitively productive learning experiences when DeepSeek was engaged as a resource for generative idea exploration, argument modelling and proof evaluation. In this respect, students leaning towards passive dependence or uncritical imitation of AI-generated results voiced fears about over-dependence and loss of their own cognitive skills and power; substantiating existing theoretical assumptions based on the cognitive offloading account (Risko & Gilbert, 2016) as well as qualitative surveys of student engagement with AI mediated feedback (Zhou et al., 2025). This conjunction of quantitative advantage and qualitative ambiguity serve to reinforce the idea that generative AI is neither a blessing nor a curse for higher-order thinking but its cognitive impact depends on how engaged students are with course materials, which pedagogical conditions govern such engagement, and how skillfully learners can leverage AI-generated content.

Third, this study is a significant method and contextual contribution within the literature as it is the first mixed-methods empirical study that simultaneously explores DeepSeek's impact on all three interacting domains in a single Iraqi EFL context. The sequential explanatory design featuring pre-post ANCOVA followed by thematic analysis of semi-structured interviews tells a more in-depth (quantitative + qualitative) story than what either would alone. Finally, the Iraqi context seems a rich terrain for an AI-assisted methodology in EFL; confined as many were within a fundamentally structuralist and culturally inflexible sector

dominated by rote memorisation over inquiry-based reasoning students noted substantial cognitive gains.

Pedagogical Implications

A few evidence-based implications for EFL practitioners, curriculum developers and institutional policymaking tend to follow from the findings. Second, the reflective evaluation of AI output should be theorized as a core pedagogic practice not an adjunct to other tasks by EFL teachers who use DeepSeek or similar generative AI tools. The demand from Carless and Boud (2018) that students should articulate their reasoning processes in accepting, altering or rejecting AI suggestions turns passive consumption of AI into an exercise of active metacognition and appears to be the key mechanism where engagement with AI can lead to genuine cognitive development. Second, curriculum designers might incorporate AI-critical literacy modules that positions learners to identify when an AI is hallucinating evaluate the internal logic of a model response or text, consider how stylistic and cultural fit plays into what they receive from an AI (and responsibly opt out). This is also possible with implementing these modules as it helps to mitigate the risks of dependence and skill-atrophy, things that students flagged themselves in the qualitative data. Third, institutional policymakers such as Iraqi universities — Al-Mustansiriyah University is running workshops about responsible usage of AI (University of Mustansiriyah 2026) so indulging them by generating strong evidence-based guidelines in the institution that do not only state which AI tools can be used across academia but: when to apply then effectively for enhancing cognitive engagement and damaging students learning.

Limitations and Directions for Future Research

There are some limitations in the current study, which need to be acknowledged. The intervention is sufficiently long that we can detect statistically significant improvements on the instruments used; however, it is not sufficient to answer less-studied questions of whether cognitive gains would last once no longer receiving the training or come with a pattern of dependency only identifiable

over more extended use of an AI. To fill this gap, future studies will require longitudinal designs which follow cognitive growth and pathways of AI utilization over time (across one or several academic years). They are further complicated by the Creative Thinking Scale which measures creativity expressed in terms of attitudes (innovative orientation, intellectual courage and flexibility), not behaviour as that found in the written product. Longitudinal studies, employing both scale-based measures and analytic rubric scores of student-generated argumentative writing across the same time period should be done in order to test whether gains in attitude towards writing argument convey observable changes in originality and quality of written argumentative output. The implications are, we hope, aimed towards all those who might be interested in the teaching and learning of ESL in any capacity as the educational situation across Iraq is so variable as to render generalizability of results limiting when restricted through a single-institution sample; within-subject investigations over multiple institutions, disciplines and English proficiency levels would greatly aid strengthening this evidence base. Finally, future research should investigate the moderating effects of how the students use AI that is, whether DeepSeek was used primarily for grammar checking, discussing ideas and thoughts or feedback indicating reflections contributing to which types or categories of human-technology interfacing are most beneficial as sustainable forms (i.e.

Finally, the findings of this study demonstrated that DeepSeek is an effective learning activity, which supported the enhancement of higher-order thinking skills for Iraqi EFL learners within a responsibly designed pedagogical design. It does not enhance cognition in itself nor detract it, the impact stems from its use aided by competent application of those tools and which users are using that tool within an instructional context. As generative AI becomes more and more common in the educational landscape worldwide, we consider it is up to educators, curriculum designers and research communities alike to help us integrate “generative

AIs” into meaningful practice with high-quality evidence based on existing learning theory; grounded on a concern that technology should not compete with but augment human capacity for critical, creative and reflective thought.

References

1. Al-Obaydi, L. H., & Pikhart, M. (2025). *Iraqi EFL students' perceptions of DeepSeek as a language learning tool*. Journal of Language and Education, 11(1), 1–15. DOI 10.1177/21582440251381683
2. Basol, G., & Gencel, I.E. (2013). *Reflective Thinking Scale: A validity and reliability study*. Educational Sciences: Theory and Practice, 13(2), 941-946.
3. Braun, V., & Clarke, V. (2006). *Using thematic analysis in psychology*. Qualitative Research in Psychology, 3(2), 77–101.
4. Carless, D., & Boud, D. (2018). *The development of student feedback literacy: Enabling uptake of feedback*. Assessment & Evaluation in Higher Education, 43(8), 1315–1325.
5. Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
6. Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. D.C. Heath.
7. Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. American Philosophical Association.
8. Flavell, J. H. (1979). *Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry*. American Psychologist, 34(10), 906–911.
9. Guest, G., Bunce, A., & Johnson, L. (2006). *How many interviews are enough? An experiment with data saturation and variability*. Field Methods, 18(1), 59–82.
10. Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
11. Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.

12. 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.
13. Hidayat, T., Susilaningsih, E., & Cepi, K. (2018). *The effectiveness of enrichment test instruments design to measure students' creative thinking skills and problem solving*. *Thinking Skills and Creativity*. Advance online publication. <https://doi.org/10.1016/j.tsc.2018.02.011>
14. Jiang, L., Cui, Y., Liu, R., & Luo, F. (2025). *Evaluating the critical thinking of large language models: Insights and limitations*. *Journal of Pacific Rim Psychology*, 19.
15. Jiang, Y., Liu, Y., & Chen, X. (2025). *DeepSeek-R1 versus undergraduate students: A comparative study of critical thinking performance*. *Computers & Education: Artificial Intelligence*, 7, 100245.
16. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). *ChatGPT for language teaching and learning*. *RELC Journal*, 54(2), 537–550.
17. Moon, J. (2004). *A handbook of reflective and experiential learning: Theory and practice*. RoutledgeFalmer.
18. Paul, R., & Elder, L. (2006). *Critical thinking: The nature of critical and creative thought*. *Journal of Developmental Education*, 30(2), 34–35.
19. Risko, E. F., & Gilbert, S. J. (2016). *Cognitive offloading*. *Trends in Cognitive Sciences*, 20(9), 676–688.
20. Runco, M. A., & Acar, S. (2012). *Divergent thinking as an indicator of creative potential*. *Creativity Research Journal*, 24(1), 66–75.
21. Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
22. Sosu, E. M. (2013). *The development and psychometric validation of a Critical Thinking Disposition Scale*. *Thinking Skills and Creativity*, 9, 107–119.
23. Sternberg, R. J. (2018). *A triangular theory of creativity*. *Psychology of Aesthetics, Creativity, and the Arts*, 12(1), 50–67.

24. University of Mustansiriyah. (2026, March 10). *A workshop at the College of Basic Education on employing artificial intelligence in teaching the English language.*
25. Van Manen, M. (1977). *Linking ways of knowing with ways of being practical.* *Curriculum Inquiry*, 6(3), 205–228.
26. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.
27. Wood, D., Bruner, J. S., & Ross, G. (1976). *The role of tutoring in problem solving.* *Journal of Child Psychology and Psychiatry*, 17(2), 89–100.
28. Xiao, Y., & Zhi, Y. (2023). *An exploratory study of EFL learners' perceptions of ChatGPT-assisted language learning.* *Languages*, 8(4), 212.