Exploring the Impact of LLM-Based Scaffolding on Academic Performance and the Mediating Roles of AI Literacy and Prior Knowledge

Anonymous authors

Abstract. Large Language Models (LLMs) have rapidly gained popularity among university students for tasks such as essay writing, virtual tutoring, and coding exercises. Although the number of studies applying LLM-based solutions in educational contexts has grown significantly, researchers emphasize the need for empirical studies to evaluate the effectiveness of these tools to support teaching and learning. To address this gap, this paper presents a study analyzing the effects of an LLM-based system designed to support students' self-reflection. The study compares the grades and course outcomes of two groups of students enrolled in the same Thermodynamics course during two academic years (2023 and 2024, 234 students in total), but only the 2024 cohort had access to an LLM-based chatbot specifically designed to support self-reflection. The data showed no significant correlation between students' AI literacy profiles, their prior experience with generative AI, and the adoption of the tool. However, an analysis of engagement with the tool in the 2024 cohort revealed that students who interacted more extensively with the chatbot—particularly medium and high achievers (based on prior academic performance)—demonstrated a significant improvement in their final grades.

Keywords: Learning Analytics, Large Language Models, Self-reflection, Quasi-experiment, Higher Education

1 Introduction

Large Language Models (LLMs), such as ChatGPT, have seen rapid adoption among university students [1]. Students are increasingly leveraging these tools to compose essays and address programming tasks. However, the extent to which such tools contribute to the learning process remains a topic of considerable debate [2]. Advocates highlight the educational affordances of generative AI, while detractors raise concerns about its potential misuse, often framing it as a tool for academic dishonesty. These divergent perspectives have resulted in polarized views. Nonetheless, the proliferation of generative AI across various domains appears irreversible. It is therefore upon the academic community to critically examine how students are engaging with these technologies and to assess their implications for teaching and learning.

Several studies have emerged in the past year examining student engagement with LLM-based agents in diverse educational settings. This work primarily analyzes interaction patterns—such as turn-taking dynamics and word counts—to quantify how learners communicate with AI [3], [4]. While such research provides preliminary insights into human-AI interaction, recent findings also highlight risks of cognitive offloading, where overreliance on AI tools may erode essential student competencies, including critical thinking [5]. In response, scholars have shifted focus toward leveraging LLMs to foster metacognitive skills and structured self-reflection [6, 7], with early results demonstrating promise.

Despite these advances, critical questions remain unresolved. Prior research identifies moderating factors—such as AI literacy and prior knowledge [6], [7] —that may influence the efficacy of LLM-based interventions, yet their impact on metacognitive outcomes remains underexplored. Still, empirical evidence is scarce on whether LLM-supported self-reflection activities yield superior metacognitive outcomes compared to traditional (non-LLM) approaches. These gaps underscore the need for rigorous, comparative studies to evaluate the effectiveness of LLMs as metacognitive scaffolds while accounting for learner-specific variables.

To address this need, this paper presents a quasi-experiment examining the effects of integrating an LLM-based tool into a Thermodynamics course to support student self-reflection over five weeks (*n*=131). Academic performance was compared with those of students from the same course 2023 and 2024 academic years. Data collection included: (a) student-LLM conversational logs, (b) course grades, (c) reflection activity grades, (d) a questionnaire capturing students' prior knowledge in a mid-term exam, and (e) a self-reported questionnaire on AI literacy, and (f) perceived utility of the tool (for cohort 2024) and prior knowledge on the course topics. These datasets were cross-analyzed to address two **research questions**:

- **RQ1**. To what extent does the LLM-based scaffolding tool enhance students' academic performance, measured by final course and reflection activity grades?
- **RQ2**. How do variations in students' (a) AI Literacy and (b) Prior Knowledge of the subject mediate their engagement patterns with the tool?

2 State of the art

2.1. Supporting self-reflection with LLMs

Self-reflection constitutes a fundamental component of metacognitive processes, enabling learners to identify misconceptions, refine their skills, and critically evaluate their learning outcomes [8]. The pedagogical effectiveness of self-reflection is further amplified when supported by well-designed prompts and methodologies, which foster a more profound and adaptable application of knowledge across diverse learning contexts.

The integration of large language models (LLMs) into educational settings has opened new avenues for enhancing and scaling self-reflective practices. Unlike human tutors, LLMs can provide instant, personalized reflection support at scale, making them uniquely positioned to address accessibility challenges in education. Emerging empirical evidence suggests that when LLMs are carefully engineered with pedagogical prompts, they can effectively guide students through structured reflection, thereby supporting metacognitive development and improving learning outcomes. For instance, [9] conducted a comparative study examining LLM-guided self-reflection against traditional methods, such as questionnaire-based reflection and passive review of lecture slides. Their findings indicated that both LLM-assisted (*d* =

0.42) and questionnaire-based approaches (d = *0.39*) yielded statistically significant improvements in subsequent assessment performance compared to passive review methods (d = *0.12*). However, no significant differences were observed between the two active reflection techniques. In a complementary study, [10] demonstrated, albeit with a small sample size (n=19), that strategically designed prompts could enable LLMs to achieve an 82% inter-rater reliability with human tutors when assessing reflective outputs, highlighting their potential as adaptive tools for both guidance and evaluation. These studies collectively underscore the potential of LLMs to serve as scalable and effective scaffolds for self-reflection when aligned with sound pedagogical principles.

Despite these encouraging findings, critical gaps remain in the current literature. While existing studies, such as those by [9], suggest that LLM-supported reflection can be as effective as traditional methods, little research is directly comparing the metacognitive gains achieved through LLM-mediated reflection versus non-LLM approaches. Such comparative analyses are essential to determine whether LLMs offer unique advantages or merely replicate the benefits of established reflective practices. Second, the real-world implications of LLM-supported reflection—including potential risks such as over-reliance on AI tools —remain underexplored. A recent review of AI's role in metacognition and critical thinking highlights the urgent need for empirical studies investigating these dynamics in authentic educational settings [8]. All this underscores the necessity for further research to elucidate the conditions under which LLM-supported reflection is most beneficial, as well as the potential pitfalls associated with its use.

2.2. Factors influencing student-LLMs interactions engagement

AI literacy has been diversely conceptualized in the literature, with definitions ranging from techno-centric [11] to human-centered perspectives [12]. The framework proposed by [13] has emerged as particularly influential, defining AI literacy as the competencies that enable individuals to evaluate AI technologies critically, collaborate with these technologies effectively, and use them as a tool in any context (online, home, and at the workplace). This comprehensive conceptualization is especially pertinent for examining student-LLM interactions, as it emphasizes the dual capacity to leverage AI's potential while maintaining a critical awareness of its limitations - a crucial balance in educational contexts where cognitive agency must be preserved.

The growing recognition of AI literacy's importance has led to the development of various instruments for measuring these skills, in addition to empirical studies examining its influence on LLM usage patterns. [6] have contributed to this effort through their Generative AI Literacy Assessment Test (GLAT), a 20-item multiple-choice instrument that demonstrated AI literacy levels significantly predict student performance in GenAI-supported tasks. Complementary research by [14] reveals how AI literacy intersects with broader digital competencies and attitudes toward technology, while [11] developed a multidimensional assessment framework for secondary students that incorporates cognitive factors (knowledge and understanding), affective components (intrinsic motivation, self-efficacy, and anxiety), and behavioral dimensions (engagement and usage intention). In a similar line, [15] identified anxiety, perceived usefulness, and positive attitudes toward AI as key psychological factors influencing students' intention to engage with AI technologies. These studies show that AI literacy operates through interactions between technical understanding, emotional dispositions, and practical engagement patterns that could influence student-LLMs interactions.

While existing research has established clear relationships between AI literacy and general LLM usage, significant gaps remain regarding how these factors operate when these tools are used to support self-reflection activities. This oversight is particularly critical because self-reflection tasks inherently require sustained cognitive effort—a process likely mediated by AI literacy dimensions [11]. Furthermore, AI literacy skills should also be examined in relation to prior domain knowledge, given its established importance in learning outcomes. Our study will take this prior knowledge into account and investigate how AI literacy profiles, in conjunction with subject matter expertise, shape the effectiveness of LLM-supported metacognitive activities in authentic educational settings.

3 Study

3.1. Study context and design

This study employed a quasi-experimental design to examine the effects of an LLM-supported self-reflection intervention in an undergraduate Thermodynamics course. The course forms part of the second-year curriculum in Aerospace Science Curriculum, spanning 5 weeks with weekly sessions comprising 2 hours of theoretical instruction and 2 hours of practical exercises. The same instructor was responsible for all teaching materials and delivery in the two years in which this study was conducted (2023 and 2024). In both years, students completed mandatory self-reflection activities through the institutional Moodle platform each week (five activities in total). These activities require students to identify and summarize 2-5 key concepts learned each week and specify those areas or concepts in which they still have doubts. While these reflections do not contribute to final grades, submissions were not compulsory except for weeks 4 and 5. That's why, for this study, we focused specifically on data from these weeks.

The research compared two student cohorts: a **control** group from the 2023 offering (n=110), who completed traditional reflection activities, and an **intervention** group from 2024 (n=103), who had access to a custom OpenAI-based LLM web-based tool designed to scaffold their reflective practice. The tool, introduced in the first week of the course, was optional for students to use when preparing their weekly summaries. The instructor configured the tool with specific pedagogical constraints: it was programmed to avoid providing direct answers, instead generating questions to prompt deeper reflection aligned with weekly learning objectives. The system incorporated course materials and topic lists provided by the instructor, transforming these into structured prompts that maintained a supportive yet focused interaction style. Students accessed these guided reflection sessions through a dedicated URL for each weekly activity.

This study ensures methodological rigor while preserving validity in real educational contexts. Both cohorts received identical core course content and materials, were taught by the same instructor, and completed examinations of comparable formats designed to assess the same learning objectives. This quasi-experimental approach allowed for a meaningful comparison between traditional and LLM-supported reflection methods within an authentic educational context.

3.2. Data gathering methods

The data collected for this study are summarized in Table 1, with all anonymized datasets available upon request. Participation required informed consent, and only data from consenting students were included in the analysis.

Two self-reported instruments were employed: (1) the AI Literacy Questionnaire for defining students' AI literacy Skills, (2) the Perceived Utility Questionnaire, for capturing students' perception about the usefulness of the tool for the activity. The AI Literacy Questionnaire was adapted from validated measures targeting dimensions relevant to LLM-supported self-reflection: AI awareness (Wang et al., 2023), behavioral engagement with AI (Ng et al., 2023), and affective factors including anxiety, perceived usefulness, and attitudes toward AI (Chai et al., 2020). An exploratory factor analysis revealed three well-defined constructs: AI Anxiety (apprehension about AI use), AI Attitude (positive disposition toward AI), and AI Usefulness (perceived utility of AI).

The Usefulness Questionnaire, co-designed with the instructor to address specific pedagogical interests, comprised five 7-point Likert scale items (1 = strongly disagree, 7 = strongly agree): (1) "Using the tool was engaging"; (2) "The tool helped me master course concepts"; (3) "I was absorbed in using the tool"; (4) "The tool was essential for understanding course content"; and (5) "I would recommend this tool to other students."

Complete instrumentation details, including the full item set and validation statistics, are available at [https://shorturl.at/TKWKq].

Table 1. Data gathering techniques

Name and Label	Description
Prior Knowledge	Students' grade in an exam with questions about the main concepts of the course taken on the **third week** for evaluating prior knowledge on the course.
Final grade,	Students' final grades in the course.
Self-reflection activity Grades,	Self-reflection activity Grades for week 4 (W4) and week 5 (W5) graded in a 0-3 rubric score, adapted from [16] and normalized in a 0-1 scale.
Perceived Utility	Questionnaire for capturing the perceived utility of the tool provided (Five 7-point likert scale questions). Students filled in the questionnaire in W2.
AI Literacy	Questionnaire composed by 17 questions directly extracted from existing questionnaires organized into 5 items: AI Awareness (3 questions); AI Behavioral Intention for working with AI (3 questions); AI Anxiety (4 questions); AI Usefulness (4 questions); AI Attitude (3 questions) define the students' AI literacy profile.
Conversation Traces	Trace data about students' conversations with the GenAI-based chatbot.

3.3. Data gathering methods

To **examine the effects of the LLM-based tool on students' final grades and summary grades (RQ1)**, we performed a two-phase analytical approach.

Phase 1. We employed a comparative analytical approach to evaluate differences in **students' final grades and summary grades between cohorts**. First, we normalized course final grades using *z*-score transformation to enable fair comparison across years, accounting for potential variations in assessment difficulty (Figure 1). Before comparing performance between cohorts, we established baseline equivalence by examining **Prior Knowledge** grade distributions (Figure 2). The null hypothesis stated no significant difference in Prior Knowledge performance means between cohorts ($\alpha = 0.05$). Using Prior Knowledge scores as a stratification variable, we categorized students into four performance quartiles: low performers (q0-q25), medium-low performers (q25-q50), medium-high performers (q50-q75), and high performers (q75-q100). This stratification enabled subgroup analyses while controlling for prior knowledge. Then, for each performance quartile, we conducted between-cohort comparisons of final grades and summary grades using independent sample *t*-tests. Effect sizes were calculated using Cohen's *d* to quantify the magnitude of observed differences. All analyses were performed using Scipy [17], with statistical significance set at $p < 0.05$.

Phase 2. We examined **how engagement with the LLM tool influenced students' final and summary grades in the 2024 cohort**. We operationalized engagement through two behavioral indicators: (1) the number of activities students engaged with the tool, since it wasn't mandatory; and (2) the number of messages exchanged with the tool.

First, we calculated, for each quartile, the number of activities they engaged with and the number of messages. Second, we used the number of activities students engaged in, and the total number of messages for categorizing students into low (n=52) and high engagement (n=51) groups using the following criteria: more/less than 10 messages exchanged and more/less than 3 activities performed. This classification yielded balanced groups for comparison (52 **low-engaged students** labelled with low engagement, and 51 **high-engaged**). Finally, we run comparisons of students' final grades between the two engagement groups using Propensity score Matching (*k=1*) controlling by Prior Knowledge grades. We used the Causal Inference library in Python[1] and calculating the Average Treatment Effect (ATE). We considered the highly-engaged students as the treatment group while the low-engaged students were the control group.

To examine how students' engagement level and course outcomes vary based on students' levels of AI literacy (hereinafter AI Literacy profiles) and prior experience using GenAI (RQ2), we conducted the following analysis**.** First, we calculated the average values for questions related to Perceived Utility, AI Usefulness, AI Attitude and AI Anxiety. Then, we calculated a correlation matrix between these variables against the two behavioral engagement indicators defined (number of

[1] https://github.com/laurencium/causalinference

activities and number of messages), the students' summary grades in W4 and W5, and students' final grades.

Second, we clustered students by the AI literacy indicators (AI Usefulness, AI Attitude and AI anxiety) and analyzed their characteristics according to their Perceived Utility. Only students who have filled at least one of the questionnaires were considered (*n* = 93) and missing values were imputed by the kNN (*k* = 5) method. The number of clusters K, was defined through Elbow analysis, indicating *k* = 3 as the balanced trade-off between model complexity and intra-cluster variance. We obtained three different clusters: C0 with 38 students, C1 with 31 and C2 with 24 students. Then, to characterize each Cluster, we analyzed their behavior using the engagement and the Perceived Utility indicators. Then, we calculated the percentage of students in each quartile within each cluster to explore whether specific clusters were associated with prior knowledge. To assess patterns of tool usage, we analyzed the proportion of high- and low-engagement students across clusters, aiming to identify whether specific clusters were more actively engaged.

4 Results

4.1. Final and Summary Grades Comparison: 2023 vs. 2024 Cohorts With and Without LLM Support (RQ1)

Concerning final and activity summary grades, data showed that the observed raw final grades showed a reduction from 2023 to 2024 ($M_{(2023)}$=14.47 vs $M_{(2024)}$=10.43), but after normalization (described in Section 3.3), no significant changes were observed, across all quartiles (Supp. Table 2C). Although no significant differences were found in the average grades of students' summaries for the reflection activity in Week 4 across quartiles (Table 2), some differences emerged in Week 5 (Supp. Table 2B). Specifically, the results show a small but statistically significant difference in summary grades for Week 5 between the 2023 and 2024 cohorts for Q0 ($M_{(2023)}$ = 0.41; $M_{(2024)}$ = 0.48) and Q50 ($M_{(2023)}$ = 0.42; $M_{(2024)}$ = 0.49), with higher mean scores observed in the 2024 cohort.

Table 2: Differences between the final grades and summary grades for W4 & W5 in 2023 and 2024 cohorts

Prior Knowledge Quartile	Mean (std) 2023	Mean (std) 2024	Difference	*p*-value
Summary Grades (W4)				
q0	0.471988 (±0.076557)	0.458966 (±0.106330)	-0.013022	0.63697
q25	0.468343 (±0.059552)	0.457095 (±0.114657)	-0.011248	0.672386
q50	0.487753 (±0.039996)	0.480149 (±0.130560)	-0.007603	0.782329
q75	0.506396 (±0.041046)	0.503515 (±0.127128)	-0.002881	0.921524

Concerning the 2024 cohort's engagement with the tool, Figure 1(a) shows notable differences in students' engagement levels across quartiles. Specifically, Table

3 indicates that students in Q25 (M = 3.36 activities) and Q50 (M = 3.50 activities) demonstrated the highest levels of engagement in terms of the number of activities completed. This pattern is further supported by the number of messages exchanged with the chatbot, with students in Q25 averaging 18.80 messages and those in Q50 averaging 19.53 messages, suggesting more sustained interaction with the tool. In contrast, students in Q0 and Q70 were the least engaged, both in terms of participation in activities and messages exchanged. However, students in Q75 showed slightly higher engagement than those in Q0, particularly in the number of messages exchanged.

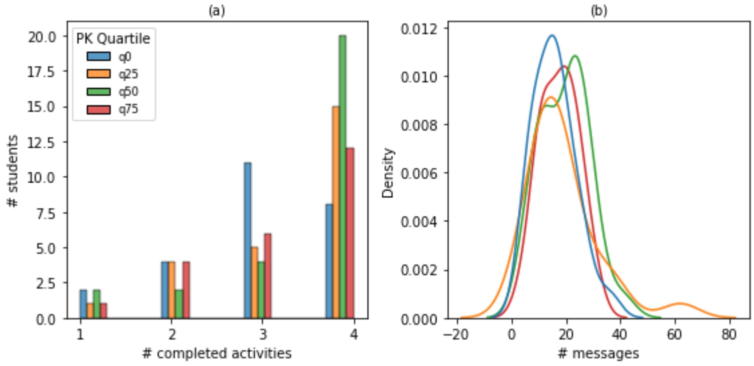


Figure 1. Level of interactions stratified by prior knowledge (PK) quartiles: (a) histogram of the number of completed activities by the students; (b) density plot of the number of messages sent by the students

Table 3. Average number of activities and messages per Prior Knowledge quartile.

Prior Knowledge Quartile	Avg Number of Activities (std)		Avg. Number of Messages (std)	
q0	3.00	(±0.912870)	15.80	(±7.686568)
q25	**3.36**	(±0.907377)	**18.80**	(±12.819256)
q50	**3.50**	(±0.922958)	**19.53**	(±8.904772)
q75	3.26	(±0.915393)	17.56	(±6.985869)

When **analyzing students' engagement with the tool in relation to their final course grades** in Figure 2, we observed notable differences. The Average Treatment Effect (ATE), controlling for prior knowledge (PK), indicates that students with high engagement achieved, on average, 2.49 points higher on the final exam than those with minimal tool usage (SE = 0.898). This effect is statistically significant, *p* = 0.006, with a 95% confidence interval ranging from 0.73 to 4.25 points. However, as illustrated in Figure 2, no significant difference was found in **summary grades** cohort when comparing high and low engagement groups. The observed average effect was an increase of only 0.012 points, with a large *p*-value (*p* = 0.718) and a 95% confidence interval ranging from −0.054 to 0.078, suggesting that this difference is likely attributable to random variation.

4.2. Relationships Between Engagement, AI Literacy, and Perceived Utility (RQ2)

For the 2024 cohort, no significant correlations were found between students' engagement levels, their AI literacy profiles, and their perceived utility of the tool (see Figure 3). An overview of cluster profiles is presented in Figure 4, which can be described as following: **Cluster 0** presents moderate levels of *AI Usefulness* (M = 3.16) and *AI Attitude* (M =3.71), and moderate to high levels of *AI Anxiety* (M=3.65), but still shows strong perceived utility of the tool (M=4.80). **Cluster 1** demonstrates moderate perceptions of *AI Usefulness* (M = 4.78) and *AI Attitude* (M = 5.55, along with relatively low levels of *AI Anxiety* (M = 2.52). Students in this cluster show medium levels of perceived utility of the tool. **Cluster 2** shows higher mean scores in both *AI Usefulness* (M = 5.47) and *AI Attitude* (M = 5.50) compared to the other clusters, but still feel anxious about using it (M=5.31). Students in this cluster reported the least perceived utility values.

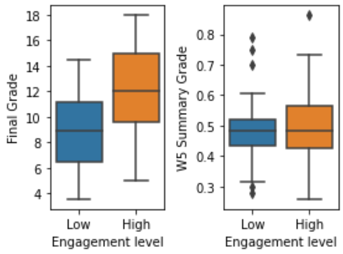
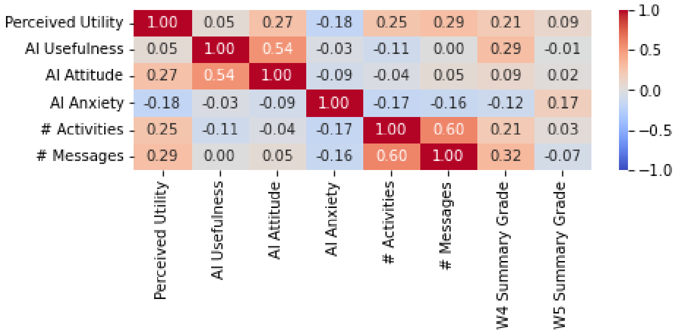


Figure 2. Academic performance measurements between high and low engagement groups.



	Perceived Utility	AI Usefulness	AI Attitude	AI Anxiety	# Activities	# Messages	W4 Summary Grade	W5 Summary Grade
Perceived Utility	1.00	0.05	0.27	-0.18	0.25	0.29	0.21	0.09
AI Usefulness	0.05	1.00	0.54	-0.03	-0.11	0.00	0.29	-0.01
AI Attitude	0.27	0.54	1.00	-0.09	-0.04	0.05	0.09	0.02
AI Anxiety	-0.18	-0.03	-0.09	1.00	-0.17	-0.16	-0.12	0.17
# Activities	0.25	-0.11	-0.04	-0.17	1.00	0.60	0.21	0.03
# Messages	0.29	0.00	0.05	-0.16	0.60	1.00	0.32	-0.07

Figure 3. AI Literacy Spearman correlation to AI tool usage.

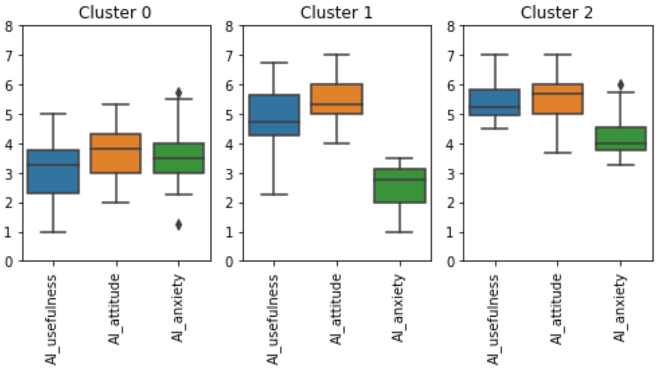


Figure 4. Box plots showing the characteristics of the cluster per the AI Literacy indicators (AI Usefulness, AI Attitude and AI Anxiety).

5 Findings

5.1 Findings on the Effects of an LLM-Based Scaffolding Tool on Students' Academic Performance (RQ1)

Our analysis revealed two findings **regarding the intervention with the tool on students' academic performance**. First, **comparing the 2023 and 2024 cohorts showed that the tool intervention did not lead to any measurable changes in students' final grades (Finding 1.1)**. This behavior is observed across the four strata defined by the quartiles of Prior Knowledge grades. Even though there was a small increase in the medium achievers (Δ_{q25}=0.06 and Δ_{q50}=0.16), we did not find this difference to be significant ($p > 0.05$). Second, **differences in summary grades were observed only in week 5, with no significant variations in the other week (Finding 1.2).** Our analysis revealed a significant positive effect on summary grades in quartiles Q0 and Q50 during week 5. One possible explanation is that, by this point, students had gained confidence in using the tool and had developed the competencies required to engage with it effectively. Still, although the increases in Q25 and Q75 were not statistically significant, the upward trend across all quartiles suggests that even limited engagement with the tool may enhance the quality of students' self-reflection work. However, there is no supporting data to confirm that these factors directly contributed to the observed increase in grades.

Concerning the **2024 cohort's engagement with the tool and its relationship to final and summary grades**, our analysis revealed 2 findings. First, the results indicate that **q25 and q50 Prior Knowledge quartiles showed more interactions with the tool on average than the others (Figure 1) (Finding 1.3)**. Second, **high-engaged students show better results in the course's final grades (Finding 1.4)**. The Average Treatment Effect (ATE), controlling by prior knowledge (PK), reveals that high-engagement students achieved, on average, 2.49 points higher on the final exam than those with minimal tool usage (*Standard Error = 0.898).* These results reinforce the finding that longer and more frequent interactions with the LLM are associated with a statistically significant and positive impact on students' overall course performance. However, given the absence of significant differences in summary grades between high- and low-engagement students within the 2024 cohort, the benefits of using AI as a learning support may manifest more clearly over the long term. This suggests that encouraging students to engage more extensively with the LLM across multiple activities and throughout different academic courses may contribute to improved final grade outcomes.

The differing results between the 2023-2024 comparison, together with the analysis of students' engagement levels within the 2024 cohort highlight two distinct dynamics: overall cohort-level effects versus individual engagement effects. One possible explanation is that the average levels of tool usage and interaction across the entire 2024 group were not sufficient to produce a measurable cohort-wide impact. This suggests that introducing an AI tool for supporting certain pedagogical activities would not be sufficient for improving academic performance for all students. The integration of these tools should be carefully prepared and pedagogically informed. Factors such as varying levels of engagement with the tool, differences in how students adapt to its use, and individual learning behaviors may affect the overall observed effects. Recent research on the design of strategies for LLM engagement reinforces that such an aspect is extremely relevant to improve learning outcomes [9].

These findings emphasize the importance of fostering meaningful engagement with AI tools to maximize their educational benefits. Encouraging strategies that promote higher interaction levels could lead to improved learning experiences and outcomes, particularly in enhancing students' metacognitive abilities, leading to improved final course grades.

5.2 AI Literacy, Prior Knowledge, and Their Influence on Student Engagement With LLM-based Scaffolded Activity (RQ2)

Regarding the **relationship between students' engagement levels with the tool and their AI literacy and perceived utility indicators**, we identified three key findings. First, **no significant correlations were observed between students' AI literacy profiles or their perceived utility of the tool and the engagement indicators (see Figure 3) (Finding 2.1)**. According to the literature, AI literacy encompasses the knowledge and skills required to interact effectively with AI technologies, including an understanding of core AI concepts, the ability to critically evaluate AI systems, and the capacity to use AI tools responsibly and ethically across various contexts (Jin et al., 2024). Although there has been an increase in the availability of instruments to assess AI literacy, these tools are predominantly based on self-reported data. As noted by Lintner (2024), self-report measures may introduce bias and fail to accurately capture an individual's actual level of literacy. This limitation may account for the lack of correlation observed between AI literacy and both perceived utility and engagement indicators in our study. Our findings suggest that the AI literacy assessment employed may not be sufficiently robust to detect students' predisposition to engage with an AI tutor within the analyzed cohort. Notice that, even if the instrument was validated, our questionnaire for capturing students' AI profile was composed of items from different questionnaires, which could have negatively impact the data collected and the results we obtained.

Second, **our analysis reveals that students in different clusters display distinct AI literacy profiles and perceptions of tool usefulness as well as engagement levels (Finding 2.2)**. We characterized students in **Cluster 0** as ***"Skeptical but Curious."*** These students reported moderate perceptions of AI usefulness and moderate to high levels of AI anxiety, yet they also expressed a generally positive perception of the AI tool used in the study. This finding suggests that despite self-reporting a certain degree of skepticism toward AI, students may still recognize and appreciate its pedagogical value when introduced in a purposeful and structured learning context.

We labeled students in **Cluster 1** as ***"Enthusiastic."*** These students self-reported strong positive attitudes toward AI and AI usefulness, along with low levels of AI anxiety. However, their reported perceived utility of the specific tool used in the study was moderate. This result suggests that these students may hold a less critical perspective on AI, viewing the use of AI-based tools for learning as expected or commonplace. As a result, their perception of the tool's utility may be tempered by familiarity rather than novelty or pedagogical impact.

We characterized students in **Cluster 2** as ***"Hopeful but Anxious**."* These students exhibited the highest scores in both AI Attitude and perceived AI Usefulness, yet also reported the highest levels of AI anxiety. Interestingly, they also reported the lowest levels of perceived utility for the tool used in the study. These findings suggest that, while these students appear to accept the integration of AI into society, their elevated anxiety may hinder their ability to fully benefit from AI-based educational interventions.

Third, **our analysis did not reveal any prevalence of a specific Prior Knowledge quartile or engagement profile within any particular cluster (Finding 2.3)**. This result diverges from prior research, which suggests that GenAI literacy has a direct influence on learners' ability to effectively engage with generative AI tools [6]. One possible explanation for this discrepancy may lie in the instrument used to assess students' AI literacy profiles. The questionnaire referred broadly to AI, without focusing specifically on generative AI in educational contexts, which may have limited its sensitivity to the skills and dispositions most relevant to engagement with the LLM-based tool.

6 Conclusions, limitations and future work

This study explored the integration of a Large Language Model (LLM)-based scaffolding tool in a Thermodynamics course, aiming to support student self-reflection over a five-week period. Through a quasi-experimental design, we investigated the impact of the tool on academic performance by comparing outcomes from two cohorts (2023 without LLM support and 2024 with LLM support). Using a multi-source dataset—including conversational logs, course and activity grades, AI literacy profiles, and perceived utility—we examined two key questions: (1) whether the LLM-based tool enhanced students' academic performance (RQ1), and (2) how students' AI literacy and prior knowledge influenced their engagement with the tool. The findings contribute to a deeper understanding of the pedagogical potential of GenAI in higher education settings.

First, findings related with the first research question offer insights into the conditions under which LLM-based scaffolding tools can influence student performance and engagement in educational contexts. While the intervention did not lead to statistically significant improvements at the cohort level, the positive association between individual engagement and academic performance highlights the importance of promoting sustained and meaningful interactions with AI tools. This supports growing evidence that the pedagogical value of GenAI depends not only on access to the technology but also on how it is embedded in instructional design and how students engage with it over time [9]. The fact that we only observed performance gains in Week 5 of the summary activity aligns with studies suggesting that students require time to develop trust, confidence, and effective strategies when interacting with GenAI systems. Moreover, the nuanced relationship between engagement and outcomes underlines the need for adaptive scaffolding and context-aware integration approaches that respond to students' prior knowledge, cognitive needs, and learning goals. Moving forward, future research should investigate how different instructional strategies, levels of guidance, and types of metacognitive prompts can optimize the educational impact of GenAI tools, while also accounting for individual learner variability and long-term effects.

Second, our findings did not reveal a direct correlation between students' AI literacy profiles and their engagement with the LLM-based tool, the clustering analysis highlights meaningful differences in how students perceive the utility of the tool based on their attitudes and anxieties toward AI. This suggests that students' subjective experiences with AI—shaped by their emotional and cognitive orientations—may influence how they interpret and value AI-based educational interventions, even if such dispositions do not translate directly into behavioral engagement. This nuanced relationship aligns with recent studies emphasizing that AI literacy is not only a matter of skills and knowledge, but also of affective and contextual factors that shape users' interaction with AI systems. Moreover, our findings underscore the importance of moving beyond generalized assessments of AI literacy to consider how specific forms of AI—such as generative AI in education—may require distinct competencies and foster varied perceptions of utility and trust [18]. Future research should consider integrating multimodal assessments and context-sensitive instruments to better capture the complex interplay between AI literacy, perception, and meaningful engagement in educational settings.

This study presents several limitations that should be acknowledged. First, although the intervention focused on a self-reflection activity, we did not employ any instrument to assess changes in students' self-reflection skills over time. This was due to both the nature of the course and the constraints typical of ecological experiments conducted in authentic classroom settings. Future research should explicitly measure self-reflection skill development to better understand how LLM-based tools can support metacognitive growth. Second, the course duration was limited to five weeks, which restricts our ability to draw conclusions about the long-term effects of the intervention. Longitudinal studies are needed to evaluate the sustained impact of LLM-based scaffolding on academic outcomes and learning behaviors. Third, at the time of the study, there were no validated instruments specifically designed to measure students' AI literacy in relation to generative AI in educational contexts. As a result, we relied on existing validated instruments that address AI literacy more generally. This may have limited the precision of our analysis and its alignment with the unique competencies required for effective GenAI engagement. Future studies should consider using more targeted, pedagogically grounded instruments to capture students' readiness and capacity to interact meaningfully with GenAI tools in learning environments.

Acknowledgments. This paper has been partially funded by the <OMITTED> and the <OMITTED>, <OMITTED>.

Declaration of Gen AI and AI-assisted technologies in the writing process. During the preparation of the paper, the authors used ChatGPT-40 and DeepSeek to check grammar and improve readability and language. After using this tool, authors reviewed and edited the content as needed and take full responsibility for the final publication.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 973–1018, Jan. 2023, doi: 10.1007/s10639-022-11177-3.

[2] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina, "Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators," *Int. J. Manag. Educ.*, vol. 21, no. 2, p. 100790, 2023.

[3] Y. Wang and X. Wang, "A study of the developmental trajectory of students' interactive dialogue model in middle school information technology course: An epistemic network analysis," *Educ. Inf. Technol.*, May 2024, doi: 10.1007/s10639-024-12760-6.

[4] H. Su, Y. Tong, X. Zhang, and Y. Fan, "Uncovering Students' Processing Tactics Towards ChatGPT's Feedback in EFL Education Using Learning Analytics," in *Blended Learning. Intelligent Computing in Education*, vol. 14797, W. W. K. Ma, C. Li, C. W. Fan, L. H. U, and A. Lu, Eds., in Lecture Notes in Computer Science, vol. 14797. , Singapore: Springer Nature Singapore, 2024, pp. 238–250. doi: 10.1007/978-981-97-4442-8_18.

[5] M. Stadler, M. Bannert, and M. Sailer, "Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry," *Comput. Hum. Behav.*, vol. 160, p. 108386, Nov. 2024, doi: 10.1016/j.chb.2024.108386.

[6] Y. Jin, R. Martinez-Maldonado, D. Gašević, and L. Yan, "GLAT: The Generative AI Literacy Assessment Test." 2024. [Online]. Available: https://arxiv.org/abs/2411.00283

[7] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing AI literacy: An exploratory review," *Comput. Educ. Artif. Intell.*, vol. 2, p. 100041, 2021.

[8] A. Goyal, "AI as a Cognitive Partner: A Systematic Review of the Influence of AI on Metacognition and Self-Reflection in Critical Thinking," *Int. J. Innov.*

Sci. Res. Technol., vol. 10, no. 3, pp. 1231–1238, 2025.

[9] H. Kumar *et al.*, "Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW2, pp. 1–30, Nov. 2024, doi: 10.1145/3687038.

[10] B. Yuan and J. Hu, "Generative AI as a Tool for Enhancing Reflective Learning in Students," *ArXiv Prepr. ArXiv241202603*, 2024.

[11] D. T. K. Ng, W. Wu, J. K. L. Leung, and S. K. W. Chu, "Artificial Intelligence (AI) literacy questionnaire with confirmatory factor analysis," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, IEEE, 2023, pp. 233–235.

[12] R. Luckin, M. Cukurova, C. Kent, and B. Du Boulay, "Empowering educators to be AI-ready," *Comput. Educ. Artif. Intell.*, vol. 3, p. 100076, 2022.

[13] D. Long and B. Magerko, "What is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, in CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–16. doi: 10.1145/3313831.3376727.

[14] B. Wang, P.-L. P. Rau, and T. Yuan, "Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale," *Behav. Inf. Technol.*, vol. 42, no. 9, pp. 1324–1337, 2023.

[15] C. S. Chai, X. Wang, and C. Xu, "An extended theory of planned behavior for the modelling of Chinese secondary school students' intention to learn artificial intelligence," *Mathematics*, vol. 8, no. 11, p. 2089, 2020.

[16] K. Miller, S. Zyto, D. Karger, J. Yoo, and E. Mazur, "Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class," *Phys Rev Phys Educ Res*, vol. 12, no. 2, p. 020143, Dec. 2016, doi: 10.1103/PhysRevPhysEducRes.12.020143.

[17] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[18] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–16.