VETTING AI for Deeper Learning: Constraining LLMs to Encourage Student Inquiry

Shan Zhang(✉), Hongming Li, Seiyon M. Lee, Noah L. Schroeder, and Anthony F. Botelho

University of Florida, Gainesville, FL 32601, USA
{zhangshan,hli3,leeseiyon,schroedern,a.botelho}@ufl.edu

Abstract. As generative AI tools like ChatGPT are integrated into education, concerns persist about over-reliance, where students seek direct answers without engaging in critical thinking. To address this, we propose VETTING, a framework for designing pedagogically-aligned applications of large language models (LLMs) in education. To explore the practical implementation of this framework, we present a case study examining the impact of a VETTING-informed chatbot that includes a verification layer designed to restrict the provision of direct problem solutions to students. In a randomized trial with 41 undergraduate students, participants interacted with either a general GPT-4o model or a VETTING-informed version designed to promote problem-solving. Findings indicate that students engaged with AI in unexpected ways, rarely asking direct questions. Instead, they sought clarification or broke problems into smaller parts. They treated the chatbot as an extension of their learning materials and rarely followed up when direct answers were unavailable. While students in the VETTING group interacted more frequently, their engagement lacked sustained inquiry. These findings highlight the complexities of designing AI tools that foster deeper learning, but offer insights into how frameworks like VETTING can be implemented to promote targeted pedagogical practices.

Keywords: Large Language Models (LLMs) · AI Literacy · AI Framework · AI Chatbot

1 Introduction

The integration of large language models (LLMs) like GPT into educational contexts presents significant opportunities to enhance student learning and engagement [18]. Educators and researchers have explored various strategies to leverage LLMs' potential while ensuring alignment with sound pedagogical principles and responsible AI practices [5]. Despite the many potential advantages LLMs can offer [1,3,5,13,15], concerns about student over-reliance on LLMs have

A. I. Cristea et al. (Eds.): AIED 2025, CCIS 2591, pp. 266–273, 2025.
https://doi.org/10.1007/978-3-031-99264-3_33

gained increasing attention as these models become more common in educational settings.

Over-reliance occurs when students depend excessively on LLM-generated responses without questioning or critically evaluating their accuracy [10]. This can lead to increased errors and a decline in independent problem-solving abilities [2]. Research has shown that students' trust in AI responses and their ability to assess the reliability of AI-generated content are key factors influencing over-reliance [8]. These are key aspects of AI literacy [11], which is a construct of growing importance as it can influence how people use and learn with AI.

Given these concerns, it is critical to help learners develop a clearer understanding of how to effectively use LLMs in learning settings. One way we can do this is by enhancing their AI literacy to foster meaningful engagement with AI tools. To mitigate over-reliance specifically, we propose the Verification and Evaluation Tool for Targeting Invalid Narrative Generation (VETTING) framework, which is designed to reshape how students interact with AI in learning environments. By implementing VETTING in a classroom setting through a case study, we specifically investigate:

RQ1. How do students using the VETTING-informed GPT model differ from those using the general GPT model in terms of academic performance, AI literacy, and engagement behaviors?

RQ2. What insights can be gained from the design and implementation of a chatbot with a verification layer to support student learning?

2 VETTING Framework

VETTING incorporates a dual-stage verification process, consisting of input pre-processing and output post-validation, to provide an iterative feedback loop rather than a binary accept or reject mechanism as is used in classification-based approaches such as Llama Guard [9]. This process is embedded within a three-layer architecture comprising the User Interaction Layer, Verification Middleware, and Language Model Integration Layer. Further details on each component are provided below.

Three-Layer Architecture. The system is structured into three layers: (A) User Interaction (UI), (B) Verification Middleware (VM), and (C) Language Model Integration (LMI). The UI layer serves as the student-facing interface, capturing queries, presenting filtered responses, maintaining session context, and applying educational scaffolding. This layer can function as any LLM-based pedagogical agent designed as a chatbot. The VM layer is the core of the framework, where input pre-processing, output validation, and post-processing occur. It detects and redirects ineffective learning strategies, such as immediate answer-seeking, while promoting productive help-seeking behaviors and scaffolded learning. The LMI layer manages prompt engineering and response generation, ensuring that outputs align with verification feedback while maintaining conversational coherence and pedagogical relevance.

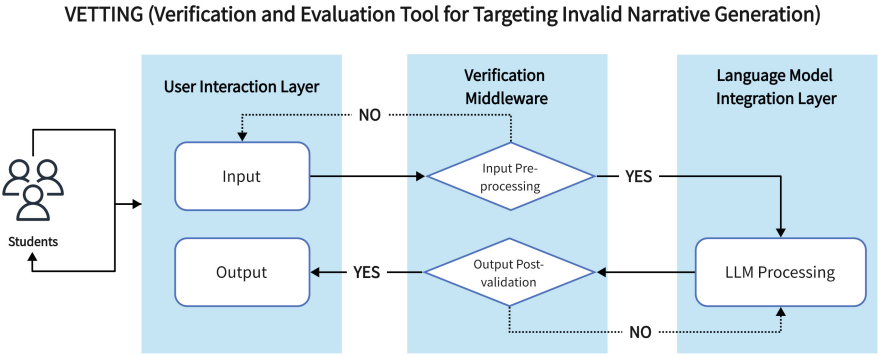

Fig. 1. The VETTING framework's three-layer architecture.

Stage 1. Input Pre-processing. This stage operates within the VM, analyzing student queries through semantic pattern recognition and contextual analysis. The filtering mechanism identifies direct answer-seeking behaviors while preserving meaningful inquiry patterns, balancing student intent with educational objectives. As illustrated in Fig. 1, VETTING processes student input to detect potentially undesirable requests. Examples include attempts to manipulate the LLM's behavior (e.g., re-prompting for specific responses) or unproductive strategies (e.g., directly requesting answers).

Stage 2. Output Post-validation Stage. This stage implements a recursive verification mechanism, where each AI-generated response undergoes multi-criteria evaluation before being delivered to the student. As shown in Fig. 1, VETTING captures the LLM-generated response before displaying it to the student, verifying its alignment with content guidelines. If the validation process detects restricted content, such as direct answers to assessment questions, it initiates an iterative refinement process, modifying the response or re-prompting the LLM until it meets content guidelines. If the response fails to meet the criteria within a set number of attempts, the system ends the refinement and informs the student that a direct answer cannot be provided and offers targeted scaffolding, such as rephrasing suggestions, problem decomposition, or related concept exploration.

3 Case Study

To examine the effect of VETTING on how students' interaction with LLM and learning, we conducted a randomized controlled study through VETTING Chat, a web-based AI chatbot building on the VETTING framework. While VETTING supports both input pre-processing and output post-validation, this paper focuses on output verification which includes Verification LLM for filtering

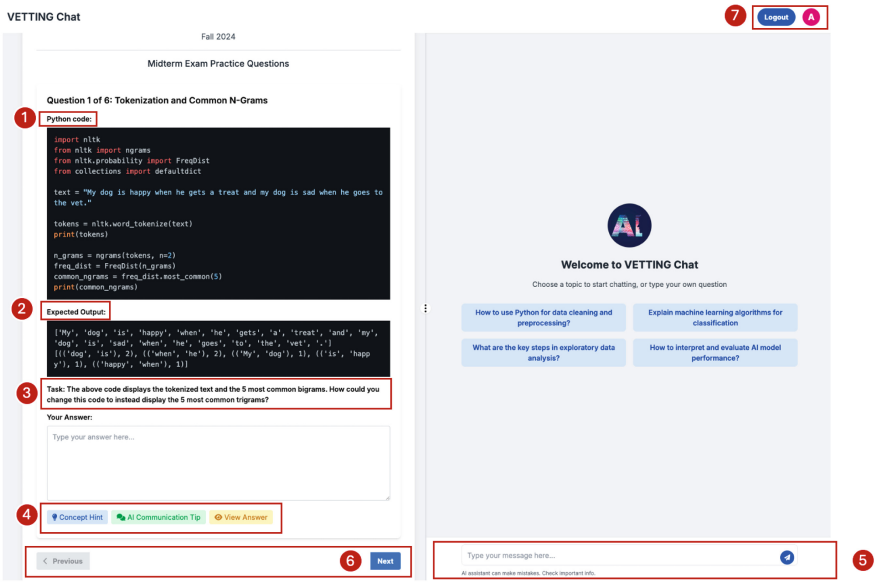

Fig. 2. A VETTING-informed Chatbot's dual-panel design with learning materials (left) and AI chat interface (right).

Chat LLM's response to students and Chat LLM for student response as part of our initial deployment stage[1].

Participants were undergraduate students enrolled in an introductory machine learning (ML) and AI course at a large southeastern U.S. university in Fall 2024. As an open course to students of any major, 70 students from 10 colleges were enrolled. Students were invited to participate in a voluntary midterm practice exam. The practice exam was administered through a web interface (Fig. 2). Forty one students participated in the study at a location of their choosing (e.g., home, school, etc.) and were randomly assigned to one of two conditions: (1) the VETTING-informed GPT-4o condition, where the model provided structured guidance without direct answers, or (2) the general GPT-4o condition[2].

The study was divided into three phases. In the first phase, participants completed six midterm practice questions that covered basic ML concepts while interacting with their assigned GPT condition. Following this, they proceeded to a post-test phase, where they answered three post-test questions without access

[1] Both models in this study used GPT-4o; detailed prompt design for both models are available on OSF: https://osf.io/fyz3e/?view_only=5cbf1b35f29e4e6cac92e503b5dad052.

[2] This study was conducted under an IRB-approved protocol and was pre-registered on OSF, which has been blinded for review.

to GPT. The study concluded with a 30-minute AI literacy test, which included 30 multiple-choice questions and one sorting question

Four types of data were collected: (1) midterm practice exam responses, (2) post-test answers, (3) an AI literacy test, and (4) backend interaction logs, including GPT conversations, query logs, response patterns, and click-stream data. The AI literacy test is adapted from existing literature [7] that researchers developed and validated items with undergraduate students similar to our context and mapped each item with an AI literacy theoretical framework [11].

A mixed-methods approach was used to analyze the data. First, non-parametric MannâĂŞWhitney U tests were conducted to examine whether there were significant differences between the VETTING and general GPT groups in post-test performance, AI literacy scores, and engagement metrics. Engagement metrics included total problem visits, total AI communication tip clicks, total concept hint clicks, and total view answer clicks. These comparisons were conducted to assess group-level differences without assuming normality, given the small sample size and data distribution characteristics. Second, two authors used an inductive thematic analysis [4] to develop a coding scheme and coded interactions between students and the chatbot for two groups.

4 Results

Not all of the participants in the study completed all of the tasks. In the VETTING group, 14 completed all tasks, with 12 finishing the AI literacy test. In the general group, 8 completed all tasks, and 6 finished the AI literacy test. High attrition was likely due to the study's optional nature and remote format.

To address RQ1, we examined whether students in the VETTING and general GPT groups differed in post-test performance, AI literacy, or engagement metrics. Mann-Whitney U tests revealed no significant differences in performance ($\text{Median}_{\text{VETTING}} = 4.5$, $\text{Median}_{\text{General}} = 4$, $U = 72$, $p = 0.278$), AI literacy Scores ($\text{Median}_{\text{VETTING}} = 20.5$, $\text{Median}_{\text{General}} = 19.5$, $U = 42.5$, $p = 0.573$), or engagement metrics: Total Concept Hint Clicks ($\text{Median}_{\text{VETTING}} = 1.5$, $\text{Median}_{\text{General}} = 1.5$, $U = 52.5$, $p = 0.833$), AI Tips ($\text{Median}_{\text{VETTING}} = 0.5$, $\text{Median}_{\text{General}} = 0.5$, $U = 56.5$, $p = 1$), Total View Answer Clicks ($\text{Median}_{\text{VETTING}} = 4.5$, $\text{Median}_{\text{General}} = 4$, $U = 57$, $p = 0.972$), and Total Problem Visits ($\text{Median}_{\text{VETTING}} = 9$, $\text{Median}_{\text{General}} = 7.5$, $U = 62.5$, $p = 0.661$).

In terms of student-LLM interaction patterns (RQ2), a total of 218 chat messages were exchanged between students. Surprisingly, only three students overall triggered an LLM response that would have contained a direct answer, requiring revision through VETTING's verification process. Upon further analysis, in the VETTING group, 12 students interacted with the chatbot, generating 42 queries, while in the general group, 8 students also asked a total of 42 questions. The full result table is available on OSF[3].

[3] Qualitative labeling is available on OSF: https://osf.io/fyz3e/?view_only=5cbf1b35f29e4e6cac92e503b5dad052.

We found that students interacted with the AI in ways that diverged from our initial expectations. They rarely used direct queries such as "Can you show me/tell me/send me?" This challenged our assumptions about the types of questions students would ask, leading to cases where the verification layer failed to detect Chat LLM responses that included direct answers. In addition, students often treated the chatbot as an extension of their learning materials, frequently copying and pasting task questions or breaking them into smaller components. When answering questions, they didn't just copy and paste what they got from the model, instead, they tended to paraphrase AI responses in their own words, indicating some critical engagement. While students in the VETTING group engaged in more, they showed limited effort in seeking alternative explanations when direct answers were unavailable.

5 Discussion

To better align the application of LLMs with pedagogical approaches, several companies have recently developed systems specifically designed to address pedagogical gaps and better support student learning. For example, recently introduced systems such as LearnLM and Claude for Education aim to promote student problem-solving ability and critical thinking, rather than relying on simple answer-giving. LearnLM incorporates pedagogical objectives into its training pipeline, combining supervised fine-tuning and reinforcement learning from human feedback to encourage models to follow instructions like "don't give away the answer" during educational dialogues [16]. Similarly, Claude's new Learning Mode applies Socratic prompting techniques and conceptual guidance to scaffold reasoning and critical thinking. However, both are early-stage efforts with limited evidence of their ability to consistently prevent direct answer-giving. While both systems incorporate instruction-tuned prompting, Socratic questioning, VETTING adds a unique post-response verification layer that explicitly blocks direct answers, providing a more robust approach to enforcing inquiry-driven interactions. In fact, this need for post-response control aligns with recommendations from Nye et al. [14] who emphasize the importance of the external validation of LLMs to ensure pedagogical alignment in dialog-based tutoring systems.

Our research offers preliminary insights into student-AI interaction patterns. It also highlighted occasional inconsistencies in LLM behavior, where both Verification LLM and Chat LLM failed to follow instructions and provided direct answers despite explicit prompts not to. This failure of LLMs to follow instructions has been documented elsewhere as well [12,19]. While we understand that instruction-following alone may not be reliable and thus requires post-response validation, the limited effectiveness of the current verification layer in handling diverse query patterns that challenge our assumptions suggests the need for more sophisticated approaches to model alignment and output control such as few-shot learning [6], multi-turn dialogue management [17] and adding input pre-processing through ML-LLM from VETTING. These findings highlight the complexities of integrating pedagogically aligned AI into learning environments

and emphasize the importance of designing tools that consider varying degrees of students' AI literacy and actively support meaningful student engagement rather than passive reliance.

Acknowledgments. We thank the National Science Foundation (#2331379), Bill and Melinda Gates Foundation, Schmidt Futures, and OpenAI for supporting this work.

References

1. Amira Roumaissa Boudjedra, H.E.K.: Promoting EFL learner's self-regulated learning through the use of artificial intelligence applications (2024)
2. Bo, J.Y., Wan, S., Anderson, A.: To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. arXiv preprint arXiv:2412.15584 (2024)
3. Chang, D.H., Lin, M., Hajian, S., Wang, Q.Q.: Educational design principles of using AI chatbot that supports self-regulated learning in education: goal setting, feedback, and personalization. Sustainability **15**(17), 12921 (2023)
4. Clarke, V., Braun, V.: Thematic analysis. J. Posit. Psychol. **12**(3), 297–298 (2017)
5. Fu, Y., Weng, Z.: Navigating the ethical terrain of AI in education: a systematic review on framing responsible human-centered AI practices. Comput. Educ. Artif. Intell. 100306 (2024)
6. Gharoun, H., Momenifar, F., Chen, F., Gandomi, A.H.: Meta-learning approaches for few-shot learning: a survey of recent advances. ACM Comput. Surv. **56**(12), 1–41 (2024)
7. Hornberger, M., Bewersdorff, A., Nerdel, C.: What do university students know about artificial intelligence? Development and validation of an AI literacy test. Comput. Educ. Artif. Intell. **5**, 100165 (2023)
8. Hunter, R., Moulange, R., Bernardi, J., Stein, M.: Monitoring human dependence on AI systems with reliance drills. arXiv preprint arXiv:2409.14055 (2024)
9. Inan, H., et al.: Llama guard: LLM-based input-output safeguard for human-AI conversations. arXiv preprint arXiv:2312.06674 (2023)
10. Korpimies, K., Laaksonen, A., Luukkainen, M.: Unrestricted use of LLMs in a software project course: student perceptions on learning and impact on course performance. In: Proceedings of the 24th Koli Calling International Conference on Computing Education Research, pp. 1–7 (2024)
11. Long, D., Magerko, B.: What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2020)
12. Lou, R., Zhang, K., Yin, W.: A comprehensive survey on instruction following. arXiv preprint arXiv:2303.10475 (2023)
13. Min, T., Lee, B., Jho, H.: Integrating generative artificial intelligence in the design of scientific inquiry for middle school students. Educ. Inf. Technol. 1–32 (2025)
14. Nye, B.D., Mee, D., Core, M.G.: Generative large language models for dialog-based tutoring: an early consideration of opportunities and concerns. In: LLM@ AIED, pp. 78–88 (2023)
15. Sahab, S., Haqbeen, J., Ito, T.: Conversational AI as a facilitator improves participant engagement and problem-solving in online discussion: sharing evidence from five cities in Afghanistan. IEICE Trans. Inf. Syst. **107**(4), 434–442 (2024)

16. Team, L., et al.: Learnlm: improving gemini for learning. arXiv preprint arXiv:2412.16429 (2024)
17. Yi, Z., Ouyang, J., Liu, Y., Liao, T., Xu, Z., Shen, Y.: A survey on recent advances in LLM-based multi-turn dialogue systems. arXiv preprint arXiv:2402.18013 (2024)
18. Zhao, R., Yunus, M.M., Rafiq, K.R.M., et al.: The impact of the use of chatgpt in enhancing student's engagement and learning outcomes in higher education: a review. Int. J. Acad. Res. Bus. Soc. Sci. **13**(12) (2023)
19. Zhou, J., et al.: Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911 (2023)