Evaluating ChatGPT-Generated Linear Algebra Formative Assessments

Nelly Rigaud Téllez[1]*, Patricia Rayón Villela[2], Roberto Blanco Bautista[3]

[1] Department of Industrial Engineering, FES Aragón, National Autonomous University of Mexico (Mexico)
[2] Universidad Internacional de La Rioja (Mexico)
[3] Department of Computer Engineering, FES Aragón, National Autonomous University of Mexico (Mexico)

Received 26 October 2023 | Accepted 23 January 2024 | Published 13 February 2024

Abstract

This research explored Large Language Models potential uses on formative assessment for mathematical problem-solving process. The study provides a conceptual analysis of feedback and how the use of these models is related in the context of formative assessment for Linear Algebra problems. Particularly, the performance of a popular model known as ChatGPT in mathematical problems fails on reasoning, proofs, model construction, among others. Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve student's achievement of intended instructional outcomes. The study analyzed and evaluated feedback provided to engineering students in their solutions, from both, instructors and ChatGPT, against fine-grained criteria of a formative feedback model that includes affective aspects. Considering preliminary outputs, and to improve performance of feedback from both agents' instructors and ChatGPT, we developed a framework for formative assessment in mathematical problem-solving using a Large Language Model (LLM). We designed a framework to generate prompts, supported by common Linear Algebra mistakes within the context of concept development and problem-solving strategies. In this framework, the instructor acts as an agent to verify tasks in a math problem assigned to students, establishing a virtuous cycle of learning of queries supported by ChatGPT. Results revealed potentialities and challenges on how to improve feedback on graduate-level math problems, by which both educators and students adapt teaching and learning strategies.

Keywords

Formative Assessment, ChatGPT, Linear Algebra, Math Word Problems, Polya's Strategy, Prompt Generator.

DOI: 10.9781/ijimai.2024.02.004

I. Introduction

Large Language Models (LLM) and the emergence of the popular ChatGPT, GPT-3.5 and GPT-4 by OpenAI [1] have spread significant developments in the context of Natural Language Processing. The underlying technology is becoming a meaningful turning point in the field of education [2].

Users enter clear commands or prompts to receive a wide range of natural-language tasks extending from text, image, videos, or code [3]. Such AI-driven educational dialogues have the potential to be a tool in education, as shown by the growing body of research, where attention focus in the improvement of active and personalized learning experience, reinforcement of learning, and assistance of the teaching processes [4] [5].

For instance, the rapid success of ChatGPT in a noticeably brief time seems to be an extremely useful tool to provide simple explanations of complex concepts [6], generate interactive educational materials like quiz questions and draft scripts for classes [7] [8]. Also, this technology can summarize longer texts [9], emphasize relevant content in a subject [10], provide learning through examples and generate formative assessment [11]. It can also improve meaningful learning by assigning writing tasks [12], generate code explanations [13], or build up critical thinking by asking students to analyze responses of ChatGPT [14].

Moreover, the use of this technology could support the generation of statistical reports with measurements of skills and knowledge [15].

Nevertheless, implementing AI-based initiatives in education requires meticulous modeling and evaluation to ensure their effectiveness in supporting academic improvement [16]. While LLM has shown its accuracy as above mentioned, when reasoning tasks engage in the realm of solving math word problems, ChatGPT may provide erroneous outputs, presentation of false information as truth in cognitive tasks [17] or causing variations in motivational or metacognitive effects [18], elicited by feedback. Consequently, the accuracy of feedback to help students could be compromised.

A. Math Word Problems

Verbal narratives, often expressed through less accurate descriptions, refer to math word problems presented in educational settings. These sorts of problems offer a comprehensive indicator of mathematical skills [19], exemplified in admissions exams designed to assess mathematical literacy.

Word problems present a realistic context described in a few sentences, where questions or dilemmas are sometimes accompanied by symbols, graphics, and pictures. Solving them requires applying mathematics [20].

* Corresponding author.
E-mail address: nerigaud@unam.mx

The relevance of math word problems has increased because they support learning over math areas, for instance, algebra, linear algebra, counting and probability, geometry, number theory or intermediate algebra.

Also, math word problems can strengthen the potential of math learning over different subjects and aim to gain experience in accordance with their organization by complex levels of thinking and reasoning through solving problems strategies [21] [22].

As math word problems usually present a textual format enriched by models and formulas, textbooks constitute a fundamental part of the teaching-learning process in the classroom, likewise, they serve as a basis for generating more balanced recommendations on the type of skills that one wishes to develop in the engineering student [23].

Given that books are a dominant educational resource that instructors review and use in teaching mathematics, these sources should facilitate opportunities for students to gain experience in problem-solving or developing new learning strategies or methods, for instance, based on common math mistakes [24][25].

In particular, and aligning with the purposes of this paper, one can use books of math word problems as a benchmark to evaluate performance of various methods. This includes examining responses when solving math word problems, considering not only accuracy, but also within the context of formative assessment [26].

B. Polya's Strategy

For the provision of thorough feedback and constructive improvement suggestions, we advocate the application of Polya's problem-solving strategy. Introduced by the distinguished mathematician George Polya, this approach comprises four key steps. These four fundamental steps can address intricate mathematical problems in a structured and systematic manner, and encompass:

- Understanding the Problem: Begin by thoroughly understanding the problem statement, identifying the knowns and unknowns, and clarifying any ambiguities.
- Find a strategy: Develop a clear and organized plan to solve the problem. This may involve drawing diagrams, breaking the problem into smaller subproblems, or considering similar problems you have encountered before.
- Execution: Implement your plan step by step, performing calculations and logical reasoning to work towards a solution.
- Looking back: Once you have a solution, review, and verify it for accuracy. Ask yourself if the answer makes sense, if it aligns with your initial understanding of the problem, and if there are alternative approaches or insights that could provide further understanding.

C. Formative Assessment

Research on formative assessment has expanded in a continuum, since Black and William [27] emphasized the need to better understand assessment for learning, as a mean to facilitate interactions between teacher, technology, and students within a learning environment that provides information for the student and teacher about the learner's performance.

Through formative assessment, and in particular by means of feedback, one could raise standards and improve learning, based on the approach of evidence, as an important opportunity to close the gap between current and desired performance by generating valuable information to both, teachers and students, consequently, yielding meaningful activities [28] [29]. Moreover, researchers have considered formative assessment as an influence on future performance [30][31].

To identify concepts involved in providing effective feedback, some authors [32] found models and characteristics of feedback, where some of the most cited authors are Hattie and Gan [33]. Additionally, Jonsson, Panadero and Lipnevich [34] [35] proposed a model, also instructional recommendations linked to different types of feedback: tasks (refers to understanding and performance when doing a task), process (the strategy needed to understand or perform a task), self-regulation (regulation of actions), and self (personal and affective aspects) [32].

Normally, teachers typically provide feedback such as comments related to the task and the self-level (personal). It is not common for them to offer comments on a solution process needed to perform the task, or at the metacognitive level (self-regulation), oriented to regulate and actively engage students' own learning [34].

More recent definitions on feedback associate tasks with information, considering it as the essence of feedback: instructors communicate it to the student with the intention of modifying his/her behavior linked to the learning. Jonsson and Panadero [34] consider as relevant components: information, gap, involved agents, and students active processing. In the latest definitions and models, Carless and Boud [36], also include similar components and oriented on how to help students to use the feedback.

From that point, Lui and Andrade [30], Panadero and Lipnevich [28], and Boud [37] posit the interaction of additional factors involved in formative assessment, which include internal process of the learner, such as motivation, and emotions elicited by feedback. These factors are related directly to behavioral response and academic achievement.

In this sense, the general model of Hattie and Gan [33] might be useful for the specific area of math word problems [38]. Despite the model of Panadero [28] requiring more research, their integrative model of feedback includes affective, motivational, and self-regulated learning processes that represent an important aspect when learning mathematics.

D. Purpose of the Study

Feedback is essential for formative assessment in the context of math word problems [38], and the intention goes toward identifying what constitutes valuable feedback, critical attributes for receptiveness and effective use of feedback supported by LLM.

It seems LLM can enhance formative assessment through machine capabilities [39][40], where some stages might occur; (a) students solve math word problems through prompts, (b) ChatGPT receives answers or queries from students (full or partial), (c) the analysis carried out by the LLM models that involves summarizing and interpretations to feedback, and adaptation, as the information oriented to adjust teaching and learning [41].

As noted, ChatGPT can provide general answers, however math problems require precision and attention, and even the most insignificant mistake can lead to incorrect answers and frustration.

Therefore, when experienced instructors identify common math mistakes, this could lead to valuable learning opportunities.

The objective is to develop a framework for formative assessment in mathematical problem-solving using LLM. This framework aims to generate prompts, supported by common Linear Algebra mistakes within the context of concept development and problem-solving strategies.

The objective of this research is to highlight the conjunction between teacher evaluations and their integration with ChatGPT during an evaluation process. We took this initiative driven by the observed underperformance of students in Linear Algebra. The study aims to leverage the combined strengths of both human teaching expertise and ChatGPT's language model capabilities, enriched by

the collective teaching experience. The underlying assumption is that through an adequate and comprehensive assessment involving both agents, teachers and ChatGPT, the student performance can be enhanced, potentially alleviating negative emotions associated with studying this subject.

We focus on examining feedback in math word problems and evaluate the potential of ChatGPT, when oriented with prompts in the process of solving mathematical problems. Two main questions are: What is the contribution of ChatGPT or the instructor in formative assessment considering its appropriate components? and is it possible to propose prompts based on a methodology that includes knowledge of common errors and formative components?

In the following sections, based on the theoretical and empirical background, also, from research questions, we present the research method and main results.

Finally, we discuss theoretical, methodological, and practical implications in the context of math learning and formative assessment supported by LLM.

II. Methods

A. Materials

To conduct this experiment, we chose the subject of Linear Algebra due to its recognition as a relevant mathematics. However, students find its learning challenging. Also, teachers find it challenging to teach.

We used a popular book of Linear Algebra named "Linear Algebra and its applications" by Lay and other authors [42] which includes a special section "Practice Problems". These problems serve to address potential challenges within the exercises or serve as a valuable prelude, and their solutions often include beneficial tips and cautions concerning homework.

We implemented a distance learning class, where students had to address, for this experiment, a set of five Linear Algebra practice problems from the specified textbook, aligning with the curriculum of a Linear Algebra course. Below, there are the five practice problems arranged from the easiest to the most difficult:

Problem one. "*Construct one different augmented matrix for linear systems whose solution set is x1=-2, x2=1, x3=0*".

Problem two. "*Suppose the solution set of a certain system of linear equations can be described as x1=5+4x3, x2=-2-7x3, with x3 free. Use vectors to describe this set as a line in R3*".

Problem three. "*Suppose a 4x7 coefficient matrix for a system of equations has 4 pivots. Is the system consistent? If the system is consistent, how many solutions are there?*"

Problem four. "*Suppose an economy has three sectors: Agriculture, Mining, and Manufacturing. Agriculture sells 5% of its output to Mining and 30% to Manufacturing and retains the rest. Mining sells 20% of its output to Agriculture and 70% to Manufacturing and retains the rest. Manufacturing sells 20% of its output to Agriculture and 30% to Mining and retains the rest. Determine the exchange table for this economy, where the columns describe how the output of each sector is exchanged among the three sectors.*"

Problem five. "*Let A be a 4 x4 matrix and let x be a vector in R4. What is the fastest way to compute A^2x? Count the multiplications.*"

These exercises included two at a basic level, two at an intermediate level, and one at an advanced level. Additionally, a concluding question addressed students' emotional responses to the learning process, encompassing emotions such as boredom, anxiety, anger, indifference, and frustration [43], which have been identified as pertinent emotional reactions to feedback in mathematical learning [29].

The process and results of each exercise, along with the emotion expressed by the learner, when applicable, were used to formulate a series of prompts. These prompts were designed to elicit feedback from the student before the instructor's review, considering both a problem-solving approach and the identification of compound emotions.

B. Participants

Our experiment took place at the Faculty of Superior Studies Aragon from the National Autonomous University of Mexico. The online classes' main goal is to improve knowledge, comprehension and problem solving of Linear Algebra.

We invited thirty-five low performance students from Industrial (60%), Mechanical (25%) and Electric-electronic (15%) careers to join the Linear Algebra course; therefore, the sample was non-probabilistic.

The total duration of the course was 32 h with four sessions per week. Three experienced teachers instructed students with explanations of Linear Algebra's fundamental concepts and resolved problems to successfully tackle the set of five Linear Algebra practice problems. Also, as requested, each of the instructors provided help to participants during interventions with ChatGPT.

Furthermore, these three teachers contributed to review and generate manual feedback to students' responses. Finally, three more teachers conducted a meta-evaluation of the feedback, as well as its comparison with ChatGPT's feedback.

The main function of ChatGPT was to provide explicit feedback according to user's prompts.

We informed all participants about the conducted experiment and obtained their consent for data collection during the process, including videotaping.

C. Tasks and Methods

As a first step, the participating students enrolled in a course of two-hour. They also engaged in assessment exercises and responded to surveys in which they provided information about their self-perception of learning difficulties. As a result of this process, information about whether the student has learned difficulties is stored in the "Common Linear Algebra mistakes" (Table I).

Table I lists a sample of a few common Linear Algebra mistakes related to concept development and problem solving. Three instructors analyzed answers. We classified outputs in accordance with Polya's strategy [22] and provided exemplifications of recommendations for students based on the prompts.

The diagram on Fig. 1, shows a general process to help the instructor to give better feedback to students based on the Polya's method, the student's emotion, and fundamental common Linear Algebra mistakes as an entrance to LLM.

An expert in the math field is necessary to obtain effective feedback, by identifying common errors of the math discipline, which are then stored in the feedback database. In this case, we focus on Linear Algebra problems and utilize the Polya's method to identify whether the error generated belongs to the comprehension (understanding the problem), planning (find a strategy), doing (execute), or revision stage (looking back). The instructor uses this information from the problem selection and its solution, and adopts a multifaceted strategy encompassing problem-solving processes, self-regulation, self-reflection, and the acknowledgment of mistakes. Within this comprehensive strategy, the instructor leverages these elements to generate prompts, seeking enriched feedback from ChatGPT to provide more insightful and constructive learning experience.

A relevant tool of LLM is the employment of natural language processing to generate prompts. Particularly in the context of this paper, establishing effective communication using LLM like ChatGPT is of great relevance to obtain clear and concrete answers.

TABLE I. Example of Fundamentals Common Linear Algebra Mistakes and Prompts Recommendations

	Understanding the Problem	Find a strategy	Execute	Looking Back
Doesn't identify what the problem is	Provide at least two different descriptions of the problem			
Erroneous selection of appropriate concepts and procedures		Can you explain me the concept of... Can you explain me the method... Why the method ... is not appropriate to solve the problem	Verify the outcome ... Test the solution through method ...	Why the method ... is appropriate to solve the problem
Doesn 't know how to communicate the solution	Express how the problem makes you feel			Wby is the solution effective? o Why doesn't the proposed solution cover what was expected? How can I interpret the problem?
Do not identify the characteristics of a system		Can you belp me to identify if the system is consistent, inconsistent, or dependent? How can you identify that a system is consistent, dependent or		

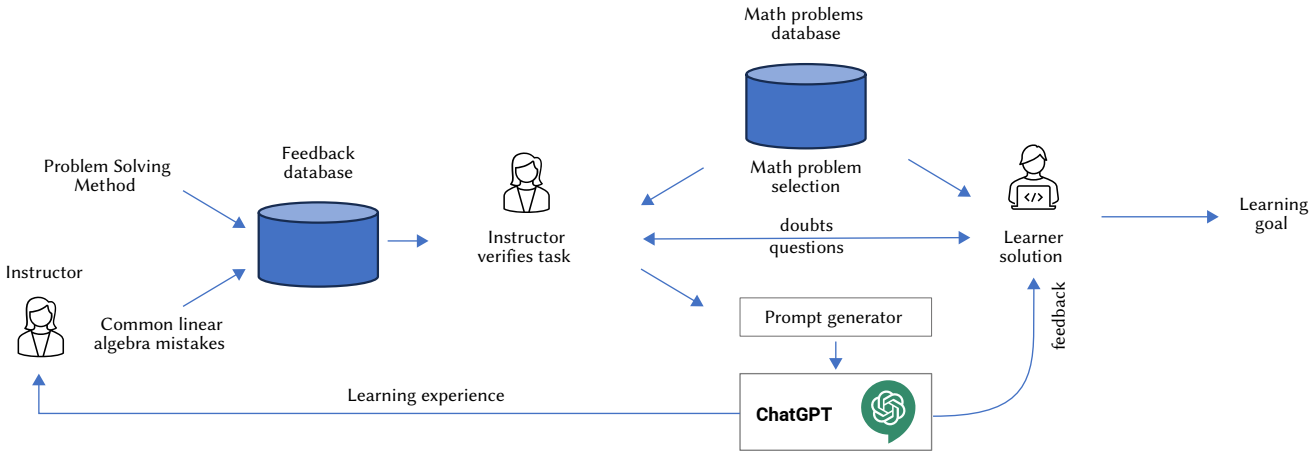


Fig. 1. Framework for formative assessment using LLM for mathematical problems.

As we have mentioned throughout the document, it is important to provide formative assessment to the student and provide some enriched prompts that the student can use with LLM.

To elicit an appropriate and constructive response from ChatGPT for students, one effective approach involves crafting specific questions. These questions serve to generate targeted feedback, incorporating motivating elements that enhance the overall quality of the student's training.

According to the prompt generation stage, the required information is:

1. Teacher Role and Course Features
2. Criterion
   i. Give the problem and the correct solution
   ii. Solving process: Polya's strategy
   iii. Solving process stage: compression, planning, doing or revision
   iv. Specify self-regulation: detail, precision, and tone
   v. Specify Self: student emotion and recommendations
   vi. Emphasize the mistake
3. Give the task (problem and solution) to ChatGPT
4. Request ChatGPT, with the information numbered as 1, 2 and 3, to generate a teaching strategy
5. Request ChatGPT to exemplify the strategy according to step 3.

For instance, generic prompts are:

Prompt 1: *I am (1) the interest is in the following math problem* (i).

Prompt 2*: For the given problem consider the* (ii) *at the phase of* (iii), *use* (iv) *for* (v).

Prompt 3: *Identify the process stage to improve...*

Prompt 4: *Request some resources...*

Some examples for prompt generation are in Fig. 2.

Prompt 1 I am a teacher of Linear Algebra for engineering bachelor the interest is in the following math problem: "Construct an augmented matrix for linear systems whose solution set is x1=-2, x2=1, x3=0", test the following "4x1+6x2+3x3=-2, -2x1+5x2+2x3=9 and x1-7x2+4x3=-9"
Prompt 2 For the given problem consider the Polya's solving process at the phase of "problem understanding". Use adequate tone and accuracy for a frustrated student.
Prompt 3 For the given problem consider the Polya's solving process at the phase of "search strategy". Use adequate tone and accuracy for a frustrated student.
Prompt 4 Give me recommendations for public link resources for the students to improve "search strategy" for "matrices"

Fig. 2. Example of prompts generation.

To examine the performance of both agents (ChatGPT and teachers) on math word problems in the context of formative assessment we based on the models of Hattie and Timperley [32] to structure relevant components on feedback, in the sense to reduce gaps between current understanding or performance and the learning goal. Furthermore, this study delves into the association of emotions triggered by feedback and self-regulation [29][34], as outlined in a well-established model for mathematical word problems. It analyzes the intricate process of solving these problems, emphasizing a thoughtful and systematic approach for complete comprehension [20]. The results are presented in the following section.

III. Results

We present the results of the analysis conducted on the feedback from ChatGPT and the teachers in Table II. As observed, we transformed the model components of Hattie and Gan [32] (Task, Solving Process, Self-regulation, and Self) into a sequence of yes-no response questions, which were then used for the assessment.

As seen in Table II, in the 'Task' component, teachers outperform ChatGPT, with an 85% accuracy compared to ChatGPT's 40%. Despite ChatGPT being capable of solving all five problems when requested individually, it makes errors when reviewing solutions generated by others. For instance, one of the most frequent errors was erroneously grading an incorrect student's response as correct.

TABLE II. Evaluation of Formative Assessment on Math World Problems

Component	Feedback	Frequency of Yes Answer	
		ChatGPT	Teacher
Task	Does the agent give a correct answer?	40%	85%
Solving process	Does the agent provide elements for understanding the problem? e.g., verbal, schematic, tabular, and so on.	90%	10%
	Does the agent model the problem?	90%	5%
	Does the agent provide calculations to resolve the model?	90%	5%
	Does the agent interpret output(s)?	90%	90%
	Does the agent evaluate the solution?	90%	90%
	Does the agent communicate the whole solution?	80%	5%
Self-regulation	Does the agent show any sort of self-management? a) Awareness of own errors	No	Yes
	b) Timing of feedback	No	Yes
	c) Level of detail	No	Yes
	d) Accuracy	No	Yes
	e) Tone	No	Yes
Self	Does the agent encourage engagement/ commitment through answers?	80%	60%
	Does the agent promote self-efficacy? (recommendations)	90%	90%

In the 'Solution Process' component, we observe that there are some aspects in which ChatGPT shows better results than a human. This is because, being an automated process, it can generate longer responses tailored to each situation, including verbal elements to understand the problem, model the problem, display the procedure's calculations, and most of the time, communicate a final solution. On the other hand, human feedback was shorter (on average, three lines) and focused on determining whether the result was correct or incorrect. In the latter case, it briefly pointed out where in the procedure the student's first error occurred but did not provide an explanation of what the correct solution and procedure should be.

It is important to note that the evaluators independently analyzed the 'task' component within the 'solution process' component. For instance, 'Does the agent provide calculations to resolve the model?' is assigned 'yes' when the agent tries to include such calculations in its feedback, regardless of whether they are correct or not.

In the 'Self-regulation' component, evaluators decided that it was challenging to assess these aspects individually in each of the samples and that a global conclusion had to be drawn for the complete set of results.

The conclusion was that although ChatGPT can regulate aspects such as tone, the level of detail in the response, etc., this is done as part of the prompt generation. However, this is externally imposed regulation by a human and not self-regulation. In the case of the teachers, there was no indication of a response that was out of context in terms of tone, level of detail, etc. According to the meta-evaluators of the experiment, all the responses provided by the humans would be the responses a teacher would typically give in a classroom.

Finally, in the 'Self' component, we can observe that ChatGPT always considered the result of emotion interpretation to craft feedback and included elements to encourage engagement and promote self-efficacy. In this regard, it is notably contrasting that the teachers' responses did not exhibit elements indicating that they considered the emotional state of the student, and the feedback was focused on problem-solving.

In the same phase of the experiment, as noted when the difficulty of problems increased, frustration is the most common emotion as shown in Fig. 3.

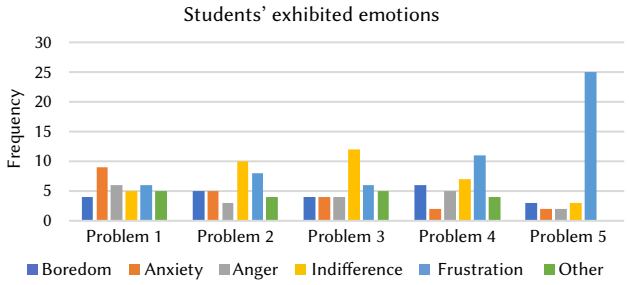


Fig. 3. Students' exhibited emotions.

Fig. 3 shows emotions exhibited by participants and provides frequencies of the experienced emotions by the students, during the solution of set of five Linear Algebra practice problems, from number one to five, as their complexity increase from less complex to more complex.

The suggested emotions include boredom, anxiety, anger, indifference, frustration, and others such as happiness or surprise. In Fig. 3, the frequency of these emotions experienced in each problem of increasing complexity is illustrated. As observed, anxiety decreased as the complexity of the problem increased. This suggests that, as students progressed in solving the problem, they were more focused on the task at hand.

The emotion of boredom remains constant throughout the problem-solving process, diminishing only in the most complex problem. The same pattern is observed for anger. Indifference increases from problem one to problem three and then decreases from problem three to problem five. In the case of frustration, it consistently increases with each new problem and experiences a significant spike in the final one. As observed, frustration appears to be the emotion that could have had the most pronounced negative impact on the group, theoretically suggesting that their performance did not improve.

Fig. 4 shows that four students did not answer any problem, and nineteen answered three problems.

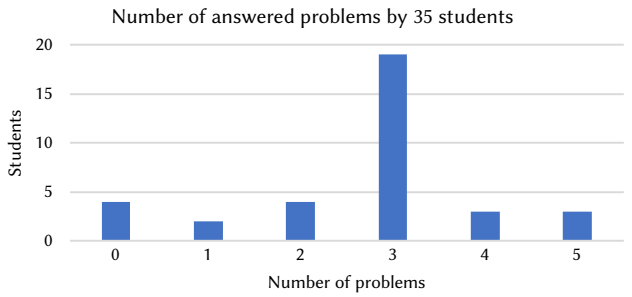


Fig. 4. Number of answered problems per participants.

Fig. 4 illustrates the distribution of students who answered a certain number of problems. The cumulative frequencies within the group of thirty-five students are as follows: four students did not solve any problem, two students answered one problem, four students answered two problems, nineteen students answered three problems, three students answered four problems, and three students successfully completed all five problems.

IV. Discussion

ChatGPT is an appropriate tool to provide more effective formative feedback due to the inclusion of four main aspects: tasks, problem solving, self- management, and self.

Interpretation of Fig. 3 and Fig. 4 suggests that there is a need for reinforcement and improvement in students' performance concerning knowledge, attitude, and dominant emotions [11][13]. This interpretation is attributed to a low proficiency in Linear Algebra, lack of comprehension of the problems, and a prominent level of distraction hindering performance improvement. Based on these findings, the recommended approach for the instructor is to prioritize feedback, incorporating both quantitative and qualitative criteria of formative assessment. Implementing strategies like Polya's problem-solving method can aid in enhancing student understanding and regaining their self-confidence.

As seen, in mathematical problems, when ChatGPT is employed independently, its performance is low. Something similar happens to the instructor. However, through the employment of the framework that includes both agents, feedback could improve learning outcomes. Additionally, the support of LLM for the students benefit their motivation to continue their math studies and reinforce math learning [26].

Nowadays, the use of these technologies is particularly important, such as LLM and the adequate use of prompts generators. Moreover, when the teachers function as a guide to construct them, recommendations are strong [44][45].

The transformative impact of LLM on mathematics learning presents key challenges that are central to the scope of this research, as follows:

To effectively integrate tools like ChatGPT into educational settings, it is imperative to establish explicit guidelines encompassing teaching and learning assessment strategies.

Specifically, within the realm of evaluation, ChatGPT tools should play a role in fostering critical thinking and logical reasoning, particularly in STEM careers, where disruptive technologies, such as those facilitated by AI, contribute to innovative and creative environments.

Considering this, prompt engineering becomes essential for shaping the approach to queries directed at ChatGPT. Well-crafted prompts should provide resources, such as relevant books available on the web, and adhere to a clear structure akin to the one proposed by the authors of this paper. This approach ensures that the generated answers are not only accurate but also engaging. The teacher's role is pivotal in this phase, serving as a verifier to confirm the correctness of the responses, as verified by the three teachers during the experiment.

For students struggling in mathematics, experiencing emotions like frustration and indifference that negatively impact their performance, ChatGPT can serve as a valuable tool. Leveraging a more human-like interaction through conversational agents, it has the potential to promote motivation and reinforce positive emotions.

V. Conclusion

Undoubtedly, the use of AI technologies with LLM represents a tool for educative support, as shown with the proposed feedback framework to improve formative assessment.

As final recommendations at the level of tasks, the instructor could propose a math word problem and assign it to the student. After the student solves it, the teacher reviews and provides regular feedback. The student asks ChatGPT to become an immersive choose-your-own task. The purpose is to reinforce the prior knowledge of the student.

For self- regulation, and from obtained feedback, students reflect and communicate about the mathematical task. Students ask ChatGPT to generate structured activities to correct his/her performance, and to encourage them to think about their learning process and math progress. Therefore, the use of ChatGPT to generate feedback is tailored to each student's needs and goals.

Another conclusion is that the teacher should encourage students to self-assess, reflect, and monitor their math work. The teacher asks ChatGPT to generate self-assessment tools, such as rubrics or the entire process for solving a math word problem that helps students evaluate their own work.

Finally, at the personal level, from provided feedback, the teacher asks ChatGPT to generate follow-up activities that encourage students to apply the feedback they have received.

For self-regulation, students engage in reflective practices based on feedback received. They utilize ChatGPT to request structured activities aimed at correcting their performance and fostering thoughtful consideration of their learning process and mathematical progress. Consequently, the use of ChatGPT for feedback generation is tailored to each student's individual needs and goals.

Alternatively, teachers can empower students to self-assess, reflect, and monitor their mathematical work. In this scenario, the teacher prompts ChatGPT to generate self-assessment tools, based on Polya's problem-solving strategy, such as rubrics or comprehensive guides for solving math word problems, facilitating students in evaluating their own work.

On a personal level, leveraging the feedback provided, the teacher can instruct ChatGPT to generate follow-up activities. These activities are designed to encourage students to apply the received feedback,

promoting a more firsthand and practical application of their learning experience, and reducing negative emotions that hinder academic performance.

This personalized approach contributes to a more comprehensive assessment tailored to individual learning needs, supported by AI technologies.

Acknowledgment

This paper has been possible thanks to the support received from The National Autonomous University of Mexico, DGAPA, PROJECT PAPIME PE112723.

References

[1] OpenAI, "ChatGPT: Optimizing language models for dialogue," Open AI 2015-2024. Accessed: Aug 13, 2023. [Online]. Available at https://openai.com/blog/chatgpt/.

[2] W.M. Lim, A. Gunasekara, J.L. Pallant, J.I. Pallant, and E. Pechenkina, "Generative AI and the future of education: Ragnarôk or reformation? A paradoxical perspective from management educators," *The International Journal of Management Education,* vol. 21, no. 2, pp. *1-13*, 2023, https://doi.org/10.1016/j.ijme.2023.100790.

[3] J. Zhou, P. Ke, X. Qiu, M. Huang, J. Zhang, "ChatGPT: potential, prospects, and limitations," *Frontiers of Information Technology & Electronic Engineering*, pp. *1-6*, 2023, https://doi.or/10.1631/FITEE.2300089.

[4] C. K. Lo, "What is the Impact of ChatGPT on Education? A rapid review of the Literature," *Education Sciences* vol. 13, no. 4, pp. 410, 2023, https://doi.org/10.3390/educsci13040410.

[5] R. Gruetzemacher and J. Whittlestone, (2022). "The transformative potential of artificial intelligence," *Futures*, vol. 135, pp. *1-11,* 2022, https://doi.org/10.1016/j.futures.2021.10288.4.

[6] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the evil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1, pp. *1-24*, 2023, https://doi.org/10.1186/s40561-023-00237-x.

[7] R. Dijkstra, Z. Genc, S. Kayal, and J. Kamps, "Reading Comprehension Quiz Generation Using Generative Pre-trained Transformers," in *4th International Workshop on Intelligent Textbooks, iTextbooks*, Durham, UK, 2022, pp. 1-14.

[8] E. Gabajiwala, P. Mehta, R. Singh, and R. Koshy. "Quiz Maker: Automatic quiz generation from text using NLP," in *Futurist trends in networks and computing technologies*, vol. 936, P.K. Singh, S.T. Wierzchoń, J. K. Chhabra, and S. Tanwar, Eds. Springer Lecture Notes in Electrical Engineering, 2022, pp. 523-533.

[9] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Ellermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, ..., and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Center for Open Science*, vol. 103, 2023, http://dx.doi.org/10.35542/osf.io/5er8f.

[10] X. Zhai, (2022), "ChatGPT user experience: Implications for education," *Social Science Research Network Electronic Journal*, vol. 18, https://doi.org/10.2139/ssrn.4312418.

[11] A. Herft, "A Teacher's Prompt Guide to ChatGPT: Aligned with 'What Works Best'," CESE NSW "What Works Best in Practice", 2023. Accessed: Aug. 15, 2023. [Online]. Available: https://usergeneratededucation.files.wordpress.com/2023/01/a-teachers-prompt-guide-to-chatgpt-aligned-with-what-works-best.pdf.

[12] A. R. Mills, "Seeing Past the Dazzle of ChatGPT," Inside Higher Education, 2024. Accessed: Jan 19, 2023. [Online]. Available: https://www.insidehighered.com/advice/2023/01/19/academics-must-collaborate-develop-guidelines-chatgpt-opinion.

[13] S. MacNeil, A. Tran, D. Mogil, S. Bernstein, E. Ross, and Z. Huang, "Generating diverse code explanations using ChatGPT-e large language model," in *Proceedings of the 2022 ACM Conference of International Computing Education Research*, New York, NY, USA, Association for Computing Machinery 2022, pp. 37-39.

[14] E.R. Mollick and L. Mollick, "Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts," The Wharton School Research Paper, 2023. Accessed: Oct. 15, 2023. [Online]. Available: https://ssrn.com/abstract=4391243 or http://dx.doi.org/10.2139/ssrn.4391243.

[15] J. F. Wu, "Effective use of machine learning to empower your research," The Campus Learn, Share, Connect, 2022. Accessed: Aug 15, 2023. [Online]. Available: https://www.timeshighereducation.com/campus/effective-use-machine-learning-empower-your-research.

[16] A. Tack and C. Piech, "The AI teacher test: Measuring the pedagogical ability of blender and GPT-e in educational dialogues," in *Proceedings of the 15th International Conference on Educational Data Mining*. Durham, UK, 2022, pp. 1-8, https://doi.org/10.48550/arXiv.2205.07540, to be published.

[17] L. M. Sánchez-Ruiz, S. Moll-López, A. Nuñez-Pérez, JA. Moraño-Fernández, and E. Vega-Fleitas, "ChatGPT Challenges Blended Learning Methodologies in Engineering Education: A Case Study in Mathematics," *Applied Sciences*, vol. 13, no. 10, 2023, https://doi.org/10.3390/app13106039.

[18] Shakarian P., Koyyalamudi A., Ngu N., and Mareedu L. (2023). "An independent evaluation of ChatGPT on Mathematical Word Problems,"

[19] A. R. Strohmaier, F. Reinhold, S. Hofer, M. Berkowitz, B. Vogel-Heuser, and K. Reiss, "Different complex word problems require different combinations of cognitive skills," *Educational Studies in Mathematics*, vol. 109, pp. 89–114, 2022, https://doi.org/10.1007/s10649-021-10079-4.

[20] L. Verschaffel, B. Greer, and E. De Corte, *Making sense of word problems*, Países Bajos: Swets & Zeitlinger, 2000.

[21] T. S. Barcelos, R. Muñoz-Soto, R. Villarroel, E. Merino, and I. F. Silveira, "Mathematics Learning through Computational Thinking Activities: A Systematic Literature Review," *Journal of Universal Computer Science*, vol. 24, no. 7, pp. 815-845, 2018.

[22] G. Polya, *Cómo plantear y resolver problemas,* Cd. México, Méx.: Editorial Trillas- Colección "Serie de Matemáticas", 1969.

[23] S. Frieder, L. Pinchetti, R. R. Griffiths, T. Salvatori, T. Lukasiewicz P. C. Peterses, A. Chevalier, and J. Berne, "Mathematical Capabilities of ChatGPT," *Neural Information Processing Systems- Datasets and Benchmarks Track*, pp. 1-37, 2023, https://doi.org/10.48550/arXiv.2301.13867.

[24] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine," *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 598-604., 2023, https://doi.org/10.1016/j.jpurol.2023.05.018.

[25] A. Tack, E. Kochmar, Z. Yuan, S. Bibauw, and C. Piech, "The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues," *in Proceedings of the 18th Workshop on innovative Use of NLP for Building Educational Applications*, Toronto, Canadian Association for Computational Linguistics, 2023, pp. 785-795, https://aclanthology.org/2023.bea-1.64.pdf.

[26] Y. Hicke, G. W. Masand, and T. Gangavarapu, "Assessing the efficacy of large language models in generating accurate teacher responses," *In Proceedings of the 18th Workshop on innovative Use of NLP for Building Educational Applications (BEA 2023*, Toronto, Canada, 2023, pp. 745-755.

[27] P. Black and D. Wiliam, "Developing the Theory of Formative Assessment," *Educational Assessment Evaluation and Accountability*, vol. 21, pp. 5–31, 2009, doi:10.1007/s11092-008-9068-5.

[28] E. Panadero and A. A. Lipnevich, "A Review of Feedback Models and Typologies: Towards an Integrative Model of Feedback Elements," *Educational Research Review*, vol. 35, 2022, doi: 10.1016/j.edurev.2021.100416.

[29] A. Ramaprasad, "On the Definition of Feedback," *Behavioral Science*, vol 28, pp. 4–13, 1983 doi:10.1002/bs.3830280103.

[30] A. M. Lui and H. L. Andrade, "Inside the Next Black Box: Examining Students' Responses to Teacher Feedback in a Formative Assessment Context," *Frontiers in Education*, vol. 7, pp. 1-14, 2022, http://dx.doi.org/10.3389/feduc.2022.751548

[31] L. Allal, "Assessment and the Co-regulation of Learning in the Classroom," *Assessment in Education: Principles, Policy &. Practices*, vol. 27, no. 4, pp. 332–349, 2019 doi:10.1080/0969594X. 2019.1609411.

[32] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112. 2007, doi:10.3102/003465430298487.

[33] J. A. C. Hattie and M. Gan, "Instruction Based on Feedback," *Handbook of Research on Learning and Instruction*, R. Mayer and P. Alexander Editors New York: Routledge), 2011.

[34] A. Jonsson and E. Panadero, "Facilitating Students' Active Engagement with Feedback," in The Cambridge Handbook of Instructional Feedback Editors, London, England: Routledge, 2018, pp. 28.

[35] A. A. Lipnevich and E. Panadero. "A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions". *Frontiers in Education*, vol. 6, 2021, doi: 10.3389/feduc.2021.720195.

[36] D. Carless and D. Boud, "The development of student feedback literacy: enabling uptake of feedback," *Assessment and Evaluation in Higher Education*, vol. 43, no. 8, pp.1315-1325, 2018.

[37] D. Boud, "Sustainable Assessment: Rethinking Assessment for the Learning Society," *Studies in Continuing Education*, vol. 22, no. 2, pp. 151–167, 2000, doi:10.1080/713695728.

[38] A. Lipnevich, F. Preckel, and S. Krumm, "Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality," *Learning and Individual Differences*, vol. 47, pp. 70–79, 2016, https://doi.org/10.1016/j.lindif.2015.12.027.

[39] B. McMurtrie, "AI and the future of undergraduate writing," The Chronicle of Higher Education, 2022. Accessed: Sept. 12, 2023. [Online]. Available: https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing.

[40] A. R. Mills. "ChatGPT just got better: What does that mean for our writing assignments?," The Chronicle of Higher Education, 2023. Accessed: March 26, 2023. [Online]. Available: https://www-chronicle-com.libproxy.library.unt.edu/article/chatgpt-just-got-better-what-does-that-mean-for-our-writing-assignments.

[41] J. Warner. "Freaking Out About ChatGPT–Part I", Inside Higher Education, 2022. Accessed: Aug. 13, 2023. [Online]. Available: https://www.insidehighered.com/blogs/just-visiting/freaking-out-about-chatgpt%E2%80%94part-i

[42] D. C. Lay, S. R. Lay, and J. J. McDonald, *Linear Algebra and its applications*, Maryland, USA: Pearson (5th Ed.), 2016.

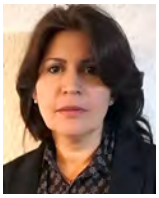
[43] A. Behera, P. Matthew, A. Keidel, P. Vangorp, H. Fang, and C. Susan, "Associating Facial Expressions and Upper-Body Gestures with Learning Tasks for Enhancing Intelligent Tutoring Systems," International *Journal of Artificial Intelligence in Education*, vol. 30, pp. 236–270, 2020, https://doi.org/10.1007/s40593-020-00195-2.

[44] F. J. García-Peñalvo and A. Vázquez-Ingelmo. "What do we mean by GenAI? A systematic literature mapping of AI-driven solutions for content generation". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4. pp. 7-16, 2023, doi: https://doi.org/10.9781/ijimai.2023.07.006

[45] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, and R. Buyya. "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19-23, 2024, https://doi.org/10.1016/j.iotcps.2023.06.002.

Nelly Rigaud Téllez

Nelly Rigaud is a Full Professor for the Industrial Engineering and Systems Department. Counselor and advisor of the Open and Distance Education System at the National Autonomous University of Mexico. She received her Engineering Doctorate from the Institute of Applied Sciences and Technology in Mexico. She holds a Master of Engineering (Planning and Projects Management) and a degree in Mechanical Engineering. Her research interests include math education and knowledge-based systems, systems modeling and simulation, and decision support systems.



Patricia Rayón Villela

Patricia Rayón has more than 20 years of experience in research and teaching in Computer Science. Experience in management and participating in research projects related to data mining, artificial intelligence, and pattern recognition issues. She is coordinator of the Master in Artificial Intelligence at UNIR-México and full professor at this university.

Roberto Blanco Bautista

Roberto Blanco received his B. Eng. Degree from the Veracruzana University. He has studies of Systems Engineering from the National Polytechnic Institute. He has more than 50 years of experience in soft computing where he has been combining teaching and counselling for many public and private organizations in soft engineering projects. His research is concerned with knowledge representation, software, and algorithms optimization.