

# Large Language Models as Problem Posers: The Case of ChatGPT, Copilot, Gemini, and Grok

*Joseph Ma. Steven Cabalo*

joseph.cabalo@student.ateneo.edu  
Ateneo de Manila University, Philippines  
Schools Division of Lipa City,  
Department of Education, Philippines

*Lester Hao*

lhao@ateneo.edu  
Ateneo de Manila University, Philippines

*Najiba Ambulo*

najiba.ambulo@deped.gov.ph  
University of Batangas, Philippines  
Schools Division of Lipa City,  
Department of Education, Philippines

*Flordeliza Ferrer*

flordeliza.ferrer@tcu.edu.ph  
Taguig City University, Philippines

*Christine Nicole Victorio*

christine.victorio@student.ateneo.edu  
National University, Philippines  
Ateneo de Manila University, Philippines

*Vanessa Zubieta*

vzubieta@slsu.edu.ph  
Southern Luzon State University,  
Philippines

**Abstract:** *This study explores the use of different large language models (LLMs) in generating new problems through Vistro-Yu's innovation techniques. A set of 30 problems were generated by each of ChatGPT, Gemini, Copilot, and Grok through structured chain-of-thought prompting. Results show that most LLMs relied on the easier techniques, often misclassified the problems it generated, and had limited diversity in the Philippine contexts it applied on the problems despite the instructions indicated in the prompt.*

## 1. Problem Posing

Problem posing has become increasingly recognized in recent years as a powerful educational strategy for cultivating students' higher-order thinking, creativity, and conceptual understanding. Defined as the generation, reformulation, or modification of problems, problem posing shifts the focus from merely solving tasks to actively constructing knowledge [1]. Evidence suggests that engaging teachers in problem posing enhances their ability to foster critical thinking, creativity, and metacognitive development in their instructional practices, particularly in mathematics and science education [2].

### 1.1. Innovation Techniques

Teachers as problem posers oftentimes face the dilemma of possibly running out of new questions or problems to pose to their students. In 2009, Vistro-Yu offered several storytelling-based innovation techniques for generating new mathematics problems based on existing problems or questions [3]. She offered six techniques adapted for this purpose: (1) replacement, where the same problem is posed but with altered quantities or units; (2) addition, which introduces a new constraint or obstacle; (3) modification, which maintains the original givens but alters the structure of the problem; (4) contextualization, which situates the problem within relevant real-world settings; (5) reversal, which reconfigures the goal and givens; and (6) reformulation, which changes the problem's type (e.g., from a proving problem to a situational one). These strategies

not only enhance the novelty and relevance of problems but also provide differentiated levels of difficulty and sophistication.

## **2. Generative Artificial Intelligence and Large Language Models**

The rapid evolution of Generative Artificial Intelligence (GenAI) has significantly transformed digital tools in education, particularly in how teachers and learners engage with content creation, such as problem posing. GenAI refers to a class of artificial intelligence models designed to generate new content, text, code, images, or even mathematical problems-based on patterns learned from massive datasets [4]. A prominent subset of GenAI is Large Language Models (LLMs), such as ChatGPT, Gemini, Copilot, and Grok, which use transformer-based architectures to understand and produce human-like text [5]. These LLMs have become increasingly relevant in educational settings due to their ability to simulate instructional dialogue, offer contextual feedback, and assist learners in constructing original problems, thus opening new avenues for problem posing [5].

### **2.1. Emergence of GenAI in Problem Posing**

Recent research indicates that GenAI is playing a growing role in reshaping how students pose and explore problems. Tools like ChatGPT and Gemini are now used as collaborative cognitive partners, assisting students in generating mathematically coherent and contextually rich problems [6]. These systems leverage vast linguistic and domain-specific data to support learners in reformulating or extending existing problems, offering alternative scenarios, and suggesting modifications that refine problem complexity and scope [7]. Moreover, educators have started adopting these AI systems as part of their pedagogical design, allowing for scaffolded interactions where learners co-construct problems with AI and receive immediate analytical feedback [8]. Studies have also demonstrated that these tools can accommodate different levels of learner ability, making the problem posing process more inclusive and differentiated [9].

### **2.2. Large Language Models (LLMs) in GenAI**

A Large Language Model (LLM) is an advanced type of AI designed to understand and generate human-like text by learning from vast amounts of language data. LLMs utilize mathematical and computational modeling for tasks like machine translation, summarization, and question answering, significantly enhancing text pre-processing capabilities [5]. LLMs are a pivotal component of GenAI, significantly enhancing natural language processing capabilities by focusing on the interaction between computer and human language [5]. However, their deployment also raises important challenges: accuracy issues, occasional generation of incorrect solutions, and limited adaptability to unique learning styles that might hinder deep conceptual understanding. LLMs often struggle with specialized queries, including high computational costs, data privacy concerns, and a lack of explainability that lead to inaccuracies in niche fields [10]. Even so, LLMs do not replace teachers; however, when used thoughtfully, they could enhance mathematical instruction, boost student motivation, and simplify teaching.

### **2.3. ChatGPT, Gemini, Copilot, and Grok**

LLMs such as ChatGPT could enhance mathematics education by offering innovative functional and pedagogical opportunities, such as problem-solving strategies similar to those found in human pre-service teachers, making it a useful demonstration tool [11]. On the other hand, Gemini, Google's GenAI, allows teachers to focus more on orchestration and personalized

intervention, using AI-generated insights for differentiated instruction [8]. Expanding the role of LLMs in education, Microsoft's Copilot has been shown to increase student engagement and understanding, with positive feedback indicating that students found the tool beneficial for their academic tasks [9]. Lastly, Grok AI, by xAI, has real-time contextual capabilities and focuses on user engagement through its freedom of expression [12].

### 3. Research Questions

In light of the potential utilization of LLMs in generating mathematical questions and problems, this study aims to evaluate how different LLMs act as problem posers when prompted to use Vistro-Yu's innovation techniques on a given base problem. In particular, the present study seeks to respond to the following questions:

- (1) Is there a consistent pattern or bias on how each large language model (LLM) applies innovation techniques?
- (2) How correctly did each LLM utilize the innovation techniques in the questions it generated?
- (3) How diverse is the Philippine contextualization as applied by each LLM in the generated questions?

### 4. Methodology

This study utilizes qualitative content analysis to investigate the issues generated by four large language models: ChatGPT (version GPT-4o), Gemini (version Gemini 2.5 Flash), Copilot (version "Quick Response"), and Grok (version Grok 3), when generating mathematical problems. A Chain-of-Thought (CoT) prompting framework, as described by Kojima et al. [13], was used to guide the model's reasoning. This technique, which improves a model's ability to solve complex tasks by prompting it through a logical, step-by-step process, was applied to generate 30 distinct mathematical problems from a single base problem. The LLM models were selected for this study due to its logged-out, free account status, which ensures that no user prompts are saved, thus preventing potential biases that could affect the quality of the model's output. Additionally, innovation techniques from Vistro-Yu [3] were incorporated into the prompt to promote the diversity of the generated problems.

This research aligns with the updated Philippine mathematics curriculum for Grade 4, focusing on students' competencies in performing addition and subtraction operations with dissimilar fractions within the Number and Algebra content domain. The foundational problem employed in this study was adapted from Tabilang et al. [14], a textbook from the Philippine Department of Education. To contextualize the problem within a Philippine setting, the scenario incorporates a culturally significant event (a birthday celebration), local and imported ingredients (Baguio cabbage and imported cabbage), and a practical situation (purchasing ingredients and preparing pancit). Accordingly, the adopted prompt used in the LLMs is as follows [15]:

Q: What are Vistro-Yu's innovation techniques on problem posing?

A: (a) replacement – posing the same problem but changing quantities, amounts, units, shapes, etc. (b) addition – posing the same problem but adding a new given or constraint or adding an obstacle. (c) modification – takes the same given but changes the problem. (d) contextualizing the problem to make it more relevant to students. (e) turning the problem around or reversing the problem – taking the same problem but taking the end goal as the given and the given as the end goal. (f) reformulation – posing the same problem in a different type (e.g. from a proving problem to a situational problem, see Butts, 1980). Depending on the innovation technique used, the new problem generated may be better, worse, or just the same in terms of the level of difficulty, sophistication, and novelty.

Q: Using innovation techniques by Vistro-Yu in diversifying problems, create one problem based on this base problem: "How can Christopher add  $\frac{1}{4}$  and  $\frac{2}{3}$ ?" Ensure that the generated problem: becomes diversified; has contexts that are Philippine-specific; is classified under the innovation techniques; and is based on the learning Competency: "The learners add and subtract dissimilar fractions."

A: Sample answer: "If Ian has  $\frac{1}{4}$  of a kalamay and Joseph has  $\frac{2}{3}$  of another kalamay, how many kalamay does Ian and Joseph have together for both of them to eat?" Innovation technique: contextualizing the problem (new problem modernizes the context of the problem by talking about particular persons engaged in eating; a slightly different problem is posed). Philippine-specific context: kalamay as a traditional Filipino snack.

Q: Using various innovation techniques by Vistro-Yu in diversifying problems, create thirty problems based on this base problem: "Alma's mother wants to cook a special *pansit* on her birthday. To make it more special she wants you to use different kinds of cabbage so she decided to go to the nearest grocery store and she saw Baguio cabbage and imported cabbage. She bought  $\frac{1}{2}$  kg of Baguio cabbage and  $\frac{1}{8}$  kg of imported Cabbage. If she used  $\frac{2}{5}$  kg of the cabbages, how many kg of cabbages were left?" Ensure that the generated problems: become diversified; have contexts that are Philippine-specific; are classified under the innovation techniques; and are based on the learning Competency: "The learners add and subtract dissimilar fractions."

A:

**Figure 1.** Structured CoT prompt used for the four LLMs on the generation of thirty problems.

Researchers (in groups of three per team) were assigned to each of Copilot, Gemini, and Grok to analyze and classify the innovation technique used on the LLM's 30 generated problems. First, each researcher independently classified the problems without knowledge of the LLM's classification nor their co-authors'. Afterwards, each team convened to decide on the innovation technique they will assign to each problem; if at least two members classified a certain problem using the same innovation technique, then the team followed that classification. Lastly, each team compared their classifications with that of the LLM's self-reported innovation technique. This prompted further qualitative discussion and analysis on the nature of the LLM's efforts to generate new problems based on a given problem.

As for ChatGPT, we refer to the results obtained from a previous study conducted by some of the authors of this paper in which they investigated ChatGPT as a problem-poser following the same methodology of the present paper [15].

## 5. Results and Discussion

In this section, we present the results per LLM and findings across the four LLMs examined. In the subsequent tables, *actual frequency* pertains to the frequency obtained from the team's categorization.

### 5.1. ChatGPT

It was found that while ChatGPT was able to utilize the context in the base problem (i.e., Philippine food), it misclassified 50% of the problems it generated [15]. Table 1 shows the number of problems per innovation technique as classified by ChatGPT and as reviewed by the researchers.

Innovation Technique	ChatGPT's Frequency	Actual Frequency
Replacement	5	18

Addition	5	0
Modification	5	0
Contextualizing	5	7
Reversing	5	5
Reformulation	5	0
Invalid	0	0
Total	30	30

**Table 1.** Classifying ChatGPT’s generated problems based on innovation techniques [15]

ChatGPT’s major classification errors fall mostly under *replacement*, which are swaps of items and fractions, but were labelled as other techniques. One instance was as follows: “A family bought  $\frac{1}{4}$  kg of shrimps from Pangasinan and  $\frac{3}{8}$  kg of clams from Batangas for their seafood stew. After cooking, they used  $\frac{5}{12}$  kg of the seafood. How much seafood remains?” ChatGPT categorized it under *addition*, but no additional constraint or obstacle was introduced.

## 5.2. Copilot

The LLM attempted to produce five problems each for the six different innovation techniques, even without being prompted to do so. After the problems have gone through review, it was found that the majority of the generated problems used the innovation technique *modification* (14 out of 30). Table 2 shows the distribution of innovation techniques as declared by Copilot and as reviewed by a team of authors. More so, Copilot only correctly classified 9 out of the 30 problems according to the innovation techniques, i.e., four for replacement, two for addition, two for modification, and one for reversing.

Innovation Technique	Copilot’s Frequency	Actual Frequency
Replacement	5	8
Addition	5	4
Modification	5	14
Contextualizing	5	0
Reversing	5	1
Reformulation	5	0
Invalid	0	3

Total	30	30
-------	----	----

**Table 2.** Classifying Copilot’s generated problems based on innovation techniques

The LLM’s attempts to reformulate the base problem produced some problems that are invalid because they are either too vague or lack the given information (3 out of 5 problems categorized by Copilot as reformulation). One example of an invalid problem is as follows: “Alma made pansit but forgot the measurements. She remembers using Bagoio and imported cabbage. Estimate their total if Bagoio was more than half.” Other problems were misclassified as reformulation, since these actually utilized the innovation technique *addition* or *modification*. To illustrate, the problem “In a situational context: the sari-sari store has  $\frac{2}{3}$  kg of cabbage left and a customer wants  $\frac{1}{2}$  kg. Is that enough?” This problem is actually modification, since it did not change from being an arithmetical problem; instead, it was the goal that was simply modified. On the other hand, there was no contextualization done because all the generated problems used the context of food and ingredients for food preparation, which was present in the base problem. Moreover, some problems were formulated as standalone (i.e., independent from the base problem), but there are problems that are continuation of the base problem. This made the classification of the problems more complicated. One such posed problem is as follows: “She noticed she mistakenly added an extra  $\frac{1}{6}$  kg of cabbage. What’s the new total used?”

The aforementioned show that Copilot seemingly still lacks the creativity and innovation skills required to generate valid mathematical problems from a given base problem. It could do replacements or minor modifications, but not contextualization nor reformulation. It might be that the contextualization did not come about because of the context being already present in the base problem.

### 5.3. Gemini

It could be gleaned from Table 3 that Gemini frequently generated word problems using replacement and addition innovation techniques. It is also worth highlighting that the team and Gemini similarly categorized word problems generated using modification, contextualization, and reversal innovation techniques. Thus, at least from this sample of 30 generated problems, Gemini correctly classified 26 problems based on the descriptions provided in the prompt.

Innovation Technique	Gemini’s Frequency	Actual Frequency
Replacement	6	9
Addition	6	3
Modification	5	5
Contextualizing	5	6
Reversing	4	4
Reformulation	4	3
Invalid	0	0

Total	30	30
-------	----	----

**Table 3.** Classifying Gemini’s generated problems based on innovation techniques

Generally, the word problems generated by Gemini showcased a wider and more diverse array of cultural contexts, such as Philippine food, practices (e.g., community spirit of *bayanihan*, tree planting), art (e.g., indigenous handicrafts), and celebrations (e.g., *fiesta*). However, considering that the intended problem solvers are Grade 4 students, the linguistic structure could be further leveled to the target set of students, especially if English is not the students’ first language. An example problem to illustrate this point: “Two barangays are sharing a water supply. Barangay A receives  $\frac{1}{2}$  of the total water, and Barangay B receives  $\frac{1}{8}$ . If  $\frac{2}{5}$  of the entire water supply is used for irrigation, and the rest is for household consumption, how much water (as a fraction of the total supply) is left for household consumption, considering both barangays' shares?” Aside from illustrating the need for improved language structuring, the same problem was originally classified by Gemini as reformulation, justifying the classification due to the problem involving reasoning and more steps. However, this problem is simply a contextualization of the given base problem, with the same quantities used and only the scenario being modified to invoke water supply and irrigation.

#### 5.4. Grok

Most of the problems generated by Grok focused on utilizing the innovation techniques of replacement and addition (6 problems each). As could be seen on Table 4, the team found that all of the classifications by Grok in terms of the innovation techniques matched theirs.

Innovation Technique	Grok’s Frequency	Actual Frequency
Replacement	6	6
Addition	6	6
Modification	5	5
Contextualizing	4	4
Reversing	5	5
Reformulation	4	4
Invalid	0	0
Total	30	30

**Table 4.** Classifying Grok’s generated problems based on innovation techniques

The problems generated by Grok generally demonstrated alignment with the given base problem, maintaining consistency in structure and context, particularly in relation to cooking

preparation and food handling. This strong adherence to the base problem may explain the perfect agreement between the team's classification and that of the LLM since it becomes easier to distinguish changes and innovations done in the problems posed by Grok.

Furthermore, attempts towards contextualization and localization were evident in the inclusion of Philippine festivals and dishes. However, some of them appear irrelevant and culturally inauthentic; these were not considered invalid because even though the underlying contexts were incorrect, Grok was still able to execute the contextualization technique by modifying the context of the base problem to another. For instance, the problem: "During a Panagbenga festival, Nanay buys  $\frac{1}{2}$  kg of Baguio cabbage and  $\frac{1}{8}$  kg of Chinese cabbage for a vegetable dish. After using  $\frac{2}{5}$  kg, how many kg are left for her to sell at the Baguio market?" illustrates this occurrence. The question demonstrates an attempt at contextualization by situating the mathematical task within the cultural setting of the Panagbenga Festival. However, the context presented is both impractical and culturally inauthentic, specifically, the scenario suggests that a leftover portion of a vegetable dish prepared for consumption would later be sold at the market. Further illustrating this are problems in which the dish and its ingredients are a mismatch. One such problem attempted to contextualize through the preparation of *laing*, an authentic dish from Bicol. However, the problem incorrectly included imported cabbage as an ingredient. Traditionally, *laing* is made using taro leaves (*gabi*), not cabbage. These instances demonstrate a lack of cultural authenticity in the problem formulation, potentially leading to confusion or misrepresentation of traditional culinary practices for the learners. It is crucial to ensure that contextualized problems accurately reflect the real-world scenarios they aim to represent.

### 5.5. Insights Across LLMs

In response to the research questions, the group found several insights in comparing the capability of each LLM to generate problems based on a given problem, along with correctly applying and classifying the innovation techniques.

First, the LLMs generally have a tendency to utilize the easier forms of innovation techniques (i.e., replacement, addition, modification). However, there are some problems in which the LLM seemingly applied more than one innovation technique. This opens the likely possibility that the innovation techniques could be combined in generating new problems. To illustrate, Grok posed the following problem: "After making pinakbet, Aling Maria has  $\frac{3}{10}$  kg of cabbage left. She bought  $\frac{1}{2}$  kg of Baguio cabbage and  $\frac{1}{8}$  kg of imported cabbage from a palengke. How many kg did she use for the pinakbet?" Aside from mentioning the wrong ingredient (i.e., cabbage in pinakbet), this question seems to involve more than one innovation technique. Grok classified this problem as reversal or turning the problem around. However, close inspection reveals that addition of a new obstacle was also introduced. In this case, the problem started with how much cabbage was left (reversal) and at the same time asking how much was used in the cooking process (addition).

Second, in terms of correctly utilizing and identifying the innovation techniques, some LLMs misclassified more frequently than others, such as ChatGPT and Copilot. However, there were instances in which the teams found it difficult to classify the generated problems because the structure of the problem entails one technique, while the details involve another. This offers the possibility of viewing innovation techniques in terms of structure and detail, therefore paving the

way for combining multiple techniques in problem generation. A problem from Grok illustrates this insight: “Alma’s mother buys  $\frac{1}{2}$  kg of Baguio cabbage and  $\frac{1}{8}$  kg of imported cabbage for a pansit bihon dish. She uses  $\frac{2}{5}$  kg but must reserve  $\frac{1}{6}$  kg for a soup. How many kg are left, and is it enough for the soup?” Structurally, the problem is a modification of the base problem because of the new question posed, while in terms of detail, there is an added constraint on reserving a portion of the soup, therefore it could also be classified as addition.

Lastly, the teams found that most LLMs have a very low tendency to deviate from the context provided in a given base problem, despite not being prompted to be limited and restricted by it. Most dwelled on Philippine food and festivities, with Gemini being the only LLM that went beyond these aspects of Philippine culture. This could serve as an avenue for further research on the training and learning aspects of GenAI and LLMs.

## 6. Conclusion

Utilizing structured CoT prompting, the researchers examined four LLMs responding as problem posers that used innovation techniques. It was found that they are able to replicate the base problem albeit using easier techniques, with context mostly not deviating from the one present in the base problem. This could serve as an impetus for teachers to exercise creativity and judgement on the variety of problems produced by LLMs, and likewise for GenAI developers to improve the learning and training of such tools.

**Acknowledgements** The present study would not have been possible without the collaboration with our fellow researchers: Mary Jane Castilla (University of Sto. Tomas), Resty Catinoy (National University and Department of Education Schools Division of Lipa City), John Patrick Cultura (La Salle College Antipolo), Bryan Ceasar Felipe (Central Luzon State University), Flordeliza Francisco (Far Eastern University), and Ma. Mina Pamela Rosario (Department of Education Schools Division of Lipa City). All of us would like to thank the Philippine Council of Mathematics Teacher Educators (MATHTED), Inc. for providing an avenue for collaboration.

## References

- [1] Cai, J., & Rott, B. (2023). On understanding mathematical problem-posing processes. *ZDM—Mathematics Education*, 56(1), 61–71.
- [2] Cai, J., Hwang, S., Jiang, C., & Silber, S. (2015). Problem-posing research in mathematics education: some answered and unanswered questions. In F. M. Singer, N. F. Ellerton, & J. Cai (Eds.), *Mathematical problem posing. From research to effective practice* (pp. 3–34). New York: Springer.
- [3] Vistro-Yu, C. P. (2009). Using innovation techniques to generate ‘new’ problems. In *Mathematical problem Solving: Yearbook 2009*, association of mathematics educators (pp. 185–207).
- [4] Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, 11(1), 22.
- [5] Mohammad, A. F., Clark, B., & Hegde, R. (2023). Large Language Model (LLM) & GPT, A Monolithic Study in Generative AI. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)* (pp. 383–388). IEEE.

- [6] Harahap, R., Simamora, Y., Lubis, N. A., Yustinaningrum, B., & Nasution, A. K. P. (2024). The role of ChatGPT in enhancing mathematics education: A systematic review. *Advances in Nonlinear Variational Inequalities*, 28(2s) (pp. 511–524)
- [7] Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers' mathematical reasoning—as an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, 7, Article 1460337.
- [8] Luzano, J. (2024). Pedagogical influence of an AI chatbot Gemini in mathematics education. *International Journal of Academic Pedagogical Research*, 8(4), 107–112.
- [9] Supriyadi, E. (n.d.). Exploring Google Bard's (Gemini) role in enhancing research articles in computational thinking and mathematics education.
- [10] Chkirbene, Z., Hamila, R., Gouissem, A., & Devrim, U. (2024). Large Language Models (LLM) in Industry: A Survey of Applications, Challenges, and Trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)* (pp. 229–234).
- [11] Harahap, R., Simamora, Y., Lubis, N. A., Yustinaningrum, B., & Nasution, A. K. P. (2024, December 10). The role of ChatGPT in enhancing mathematics education: A systematic review. *Advances in Nonlinear Variational Inequalities*, 28(2S).
- [12] <https://x.ai/>
- [13] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199–22213.
- [14] Tabilang, A. R., Arce, I. J. B., Pascua, R. V., Calayag, N. P., Dacuba, L. P., Borais, D. B., Buemia, R. B., Collao, M. T., Morandante, L. G., Danao, A. B., Nu, L. N., Gonzaga, I. A. B., & Daganta, J. A. D. (2015). *Mathematics – Grade 4: Learner's material* (First edition). Department of Education, Philippines.
- [15] Cabalo, J. M. S., Ambulo, N., Catinoy, R., Rosario, M. M. P., Hao, L., Castilla, M. J., Victorio, C. N., Calimlim Jr., R., Zubieta, V., Cultura, J. P., & Francisco, F. (2025). ChatGPT as Problem Poser: Contextualized AI-Assisted Problem Generation Using Structured Chain-of-Thought Prompting. In Kwon, O., Kaur, B., Pang, J., Noh, J., Lee, S., Han, S., Yeo, S., & Lim, M. (Eds.), *E-Proceedings of the 9th ICMI-East Asia Regional Conference on Mathematics Education (Vol. 3)* (pp. 958-962) Seoul National University, Siheung Campus, Korea: EARCOME9