



## EMPOWERING ELEMENTARY SCHOOL TEACHERS THROUGH MATHEMATICAL PROBLEM POSING WITH AI AS A REFLECTIVE FEEDBACK TOOL: A COMMUNITY SERVICE PROGRAM IN DEPOK CITY

Linda Devi Fitriana<sup>1</sup>, Meiliasari<sup>2</sup>, Puspita Sari<sup>3</sup>, Ratna Maryam<sup>4</sup>, Dessy Julita Sari<sup>5</sup>,  
Muhammad Arib Alwansyah<sup>6</sup>, Lintang Ayuza Kalila<sup>7</sup>, Zahra Audina<sup>8</sup>

<sup>1,2,3,4,5,6,7,8</sup>Department of Mathematics Education, Universitas Negeri Jakarta, Indonesia

Email : [linda.fitriana@unj.ac.id](mailto:linda.fitriana@unj.ac.id); [meiliasari@unj.ac.id](mailto:meiliasari@unj.ac.id); [puspitasari@unj.ac.id](mailto:puspitasari@unj.ac.id);  
[maryamratna@unj.ac.id](mailto:maryamratna@unj.ac.id); [j.dessysari@unj.ac.id](mailto:j.dessysari@unj.ac.id); [muhammadarib@unj.ac.id](mailto:muhammadarib@unj.ac.id);  
[lintang.ayuza@mhs.unj.ac.id](mailto:lintang.ayuza@mhs.unj.ac.id); [zahra.audina@mhs.unj.ac.id](mailto:zahra.audina@mhs.unj.ac.id)

Diterima: 1/10/2026; Direvisi: 10/10/2026; Diterbitkan: 20/10/2026

### ABSTRACT

This community service program addressed an instructional design challenge among elementary school teachers in Depok City, where the ability to construct computable arithmetic problems did not routinely extend to designing mathematically authentic and cognitively demanding tasks. The initiative aimed to assist 18 elementary school teachers in designing, evaluating, and refining mathematical tasks through a participatory problem-posing cycle using artificial intelligence (AI) as a reflective feedback partner. The qualitative descriptive evaluation examined changes across task artifacts comprising four initial group problems, collaborative group revisions, and nine fully documented individual products. Initial baseline artifacts were limited to routine single-step computations situated within superficial narrative wrappers or devoid of context. Following collaborative AI-assisted appraisal and individual refinement, the majority of documented individual tasks shifted toward multi-step or analytical and evaluative reasoning embedded in realistic-generic or situated local contexts. However, these outcomes reflect qualitative improvements in documented task artifacts rather than a directly verified gain in teacher competence, given the absence of pre- and post-tests, independent inter-rater reliability testing, or direct classroom implementation trials.

**Keywords:** *community service, mathematical problem posing, artificial intelligence, reflective feedback, mathematical reasoning, task design*

### ABSTRAK

Program pengabdian kepada masyarakat ini menangani tantangan dalam desain instruksional di kalangan guru sekolah dasar di Kota Depok, di mana kemampuan untuk menyusun soal aritmetika yang bersifat komputasional belum mencakup perancangan tugas yang secara matematis autentik dan menuntut kemampuan kognitif tinggi. Inisiatif ini bertujuan untuk mendampingi 18 guru sekolah dasar dalam merancang, mengevaluasi, dan menyempurnakan tugas-tugas matematika melalui siklus penyusunan soal secara partisipatif, dengan memanfaatkan kecerdasan buatan (AI) sebagai mitra untuk memberikan umpan balik reflektif. Evaluasi deskriptif kualitatif dilakukan terhadap perubahan pada artefak tugas, yang mencakup empat soal kelompok awal, revisi kelompok secara kolaboratif, serta sembilan produk individu yang didokumentasikan secara lengkap. Artefak awal yang menjadi tolok ukur terbatas pada soal komputasi satu langkah yang rutin, dengan narasi dangkal atau bahkan tanpa konteks sama sekali. Setelah melalui proses penilaian kolaboratif dengan bantuan AI serta penyempurnaan secara individu, sebagian besar tugas individu yang didokumentasikan berkembang menjadi



soal yang melibatkan langkah majemuk atau penalaran analitis dan evaluatif, yang ditempatkan dalam konteks realistik-generik maupun konteks lokal yang spesifik. Akan tetapi, hasil ini mencerminkan peningkatan kualitatif pada artefak tugas yang didokumentasikan, bukan peningkatan kompetensi guru yang terverifikasi secara langsung; hal ini dikarenakan tidak adanya tes awal dan akhir (pre-test dan post-test), pengujian reliabilitas antarpemilai (inter-rater reliability) secara independen, maupun uji coba penerapan langsung di dalam kelas.

**Kata kunci:** *pengabdian kepada masyarakat, penyusunan soal matematika, kecerdasan buatan, umpan balik reflektif, penalaran matematis, desain tugas*

## INTRODUCTION

Elementary mathematics tasks shape learners' perceptions of what constitutes mathematical activity, how quantitative relationships are examined, and how school mathematics intersects with authentic human experience. When classroom tasks employ context merely as a decorative narrative wrapper, students execute procedural algorithms without engaging in meaningful conceptual interpretation. Designing high-quality tasks requires elementary teachers to coordinate mathematical fidelity, cognitive reasoning demands, linguistic clarity, and situational plausibility (Glegola et al., 2026; Hatisaru, 2020; Lee, 2024; Parrish & Bryd, 2022; Thomas et al., 2024). However, baseline diagnostic observations with our teacher partners in Depok City revealed that teachers frequently defaulted to formulating conventional, single-step word problems under time constraints. These formulations rarely prompted higher-order reasoning or capitalized on meaningful local situations, exposing a substantive professional learning gap in instructional task design.

Mathematical problem posing provides an effective vehicle for professional growth because it engages educators in formulating, analyzing, and restructuring problems from rich mathematical and contextual situations. Prior scholarly inquiries indicate that problem-posing routines foster conceptual sensemaking, pedagogical reasoning, and collective ownership of mathematical tasks (Cai, 2022). Iterative task construction and collaborative critique further help teachers refine their criteria for evaluating mathematical complexity and instructional potential (Hwang et al., 2024). A comprehensive meta-analysis also confirmed that structured problem-posing interventions exert a statistically significant positive effect on task quality and task formulation competence (Zhang et al., 2024). Nevertheless, existing professional development models rarely investigate how emerging generative technologies can be systematically leveraged to support this iterative task-design cycle without eroding teacher pedagogical autonomy. Accordingly, this study examines how ChatGPT-4o can function as a scaffold within Blum's modeling cycle, enabling teachers to iteratively refine mathematical tasks while retaining control over their instructional decisions (Brunetto & Iacono, 2026; CHUI, 2026a, 2026b; Spreitzer et al., 2024; Trevisani & Maltempi, 2026).

Contextual authenticity introduces an additional instructional hurdle, as teachers often struggle to distinguish between superficial story problems and genuinely situated mathematical challenges. While educators acknowledge the pedagogical merit of realistic tasks, their authored tasks frequently feature artificial scenarios or sterile computational prompts (Chavarría-Arroyo & Albanese, 2023). This programmatic challenge is directly relevant to Depok City, where local markers—such as *Belimbing Dewa* orchards, Margonda *angkot* transit routes, and the Dian Al-Mahri Golden Dome Mosque—offer fertile ground for realistic task modeling. Simply inserting local nouns into a computation, however, fails to produce authentic inquiry unless situational constraints directly determine the mathematical relationships and

solution pathways. Elementary teachers therefore require scaffolded collaborative mechanisms to transform regional contexts into substantive drivers of mathematical reasoning.

Generative artificial intelligence (*AI*) offers dynamic possibilities for rapid task ideation, local context adaptation, and automated feedback on cognitive expectations (Kasneci et al., 2023). Contemporary international policy frameworks emphasize, however, that educational *AI* integration must preserve human agency, ethical accountability, and pedagogical oversight (Miao & Cukurova, 2024; Miao & Holmes, 2023). Empirical comparisons between algorithmic output and human feedback similarly indicate that while *AI* delivers immediate alternative formulations, teachers rely on professional judgment to establish instructional validity and trust (Henderson et al., 2025). The novelty of this community service program lies in positioning generative *AI* not as an authoritative problem author, but as a conversational feedback tool that stimulates collaborative critique, contextual auditing, and reflective revision.

Accordingly, this community service program aimed to empower elementary school teachers in Depok City to design mathematics tasks characterized by richer contextualization, elevated reasoning demands, and heightened situational authenticity. The guiding program inquiry was: How can generative *AI*, utilized as a reflective feedback tool within a participatory problem-posing cycle, support elementary school teachers in enhancing the contextual, cognitive, and authentic dimensions of their instructional tasks? The program evaluation focused on documented artifact shifts across collaborative and individual design stages. The subsequent sections outline the participatory methods, artifact analyses, design paradoxes, and practical implications for sustainable school-based professional development.

## METHODS

This community service initiative utilized a participatory design workshop model coupled with a qualitative descriptive artifact analysis. The participatory format directly addressed the community partner's expressed need for practical task-design strategies through structured cycles of collaborative drafting, algorithmic feedback analysis, critical reflection, and task revision. The participant cohort comprised 18 in-service elementary school teachers from Depok City, all holding bachelor's degrees and aged between 25 and 44 years. The training materials included *Spot It* cards as visual stimuli for spontaneous task generation, preset generative *AI* prompt templates tailored to regional Depok contexts, and standardized artifact evaluation rubrics. The overall procedural workflow that guided the participants through each collaborative intervention stage is outlined in the following diagram.

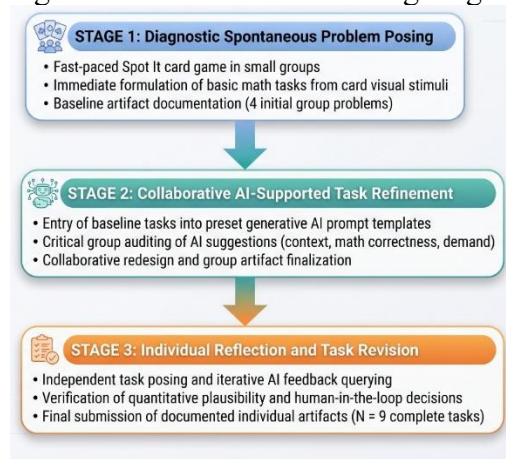


**Figure 1. Participatory Stages of the Community Service Program**



Figure 1 illustrates the three sequential stages that structured the community empowerment workshop. In Stage 1, teachers played *Spot It* in small groups of four to five participants, requiring the round loser to immediately compose a spontaneous mathematical problem based on an identified card symbol. In Stage 2, groups fed their baseline tasks into preset *AI* prompts to generate three contextualized variations, followed by small-group audits to accept, adjust, or discard the algorithmic outputs based on curricular criteria. In Stage 3, each educator independently formulated a novel task, queried the *AI* tool for reflective critique, audited the output, and finalized their instructional problem. Although ten individual artifacts were initially submitted, one submission contained incomplete documentation across design steps and was excluded, leaving nine fully documented individual products for comprehensive qualitative evaluation.

The analytical framework employed qualitative descriptive artifact analysis to assess task transformations across the intervention stages. Each instructional problem was systematically classified across three core analytical dimensions: contextualization, reasoning potential, and context authenticity (Chavarría-Arroyo & Albanese, 2023). The coding criteria and operational category definitions utilized to analyze the collected task artifacts are presented in the following table.

**Table 1. Analytical Coding Dimensions and Category Definitions for Task Artifacts**

| Dimension           | Category                     | Operational Definition                                                                                                                        |
|---------------------|------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| Contextualization   | No context                   | The problem consists purely of bare mathematical numbers or computational symbols without any situational background.                         |
|                     | Surface contextualization    | The context serves purely as a decorative narrative wrapper where mathematical quantities can be extracted without contextual interpretation. |
|                     | Meaningful contextualization | The narrative situation provides meaningful clues or supports the interpretation of mathematical procedures and quantitative results.         |
|                     | Context-driven task          | The context fundamentally shapes the problem structure, mathematical constraints, and required problem-solving strategies.                    |
| Reasoning Potential | Routine                      | The task requires straightforward, single-step algorithmic execution or direct formula application.                                           |
|                     | Multi-step                   | The task requires connecting two or more sequential computational steps to reach a correct solution.                                          |
|                     | Analytical evaluative        | / The task demands comparing alternatives, critiquing mathematical claims, or justifying decisions under given constraints.                   |
|                     | Generative modeling          | / The task requires formulating original mathematical representations, exploring open-ended scenarios, or creating models.                    |



|                             |                          |     |                                                                                                                                    |
|-----------------------------|--------------------------|-----|------------------------------------------------------------------------------------------------------------------------------------|
| <b>Context Authenticity</b> | Non-authentic artificial | /   | The scenario presents implausible quantities, unrealistic human behaviors, or highly contrived situations.                         |
|                             | Realistic yet generic    |     | The scenario is plausibly drawn from everyday life but remains generic without specific geographic or cultural grounding.          |
|                             | Realistic relevant       | and | The scenario reflects plausible real-world practices that are recognizable and meaningful to the targeted student cohort.          |
|                             | Authentic situated       | and | The scenario is deeply embedded in identifiable local physical environments, regional practices, or genuine community constraints. |

Table 1 delineates the operational benchmarks used to assess the qualitative trajectory of teacher-posed mathematical problems. The contextualization dimension examines whether narratives actively drive the mathematical inquiry or simply function as superficial packaging. The reasoning dimension evaluates the cognitive complexity demanded of students, distinguishing direct calculations from multi-stage analytical judgments. Finally, the authenticity dimension appraises whether numerical magnitudes and scenarios reflect plausible real-world dynamics. The entire analytical procedure followed four systematic steps: document sorting by design stage, rubric-based categorization, cross-stage descriptive tallying, and qualitative excerpt analysis. Because the program scope did not include independent double-coding or inter-rater reliability testing, the recorded patterns are reported strictly as descriptive indicators from this community engagement rather than generalizable causal metrics.

## RESULTS AND DISCUSSION

### Descriptive Artifact Transformation

The diagnostic baseline artifacts confirmed that elementary school teachers initially operated in a computational mode when designing tasks under temporal pressure. All four initial group problems (100.0%) were categorized as requiring routine reasoning, with three tasks exhibiting surface contextualization and one presenting no context whatsoever ( $1 + 1 = 2$ . *Is it OK?*). Following the collaborative AI-supported workshop sessions and individual refinement cycles, the final nine individual artifacts demonstrated noticeable qualitative shifts. To illustrate the proportional distribution shifts across the three analytical dimensions, the comparative coding tallies are summarized in the following table.

**Table 2. Comparative Coding Distributions Between Initial Baseline and Final Individual Task Artifacts**

| Dimension                  | Category                     | Initial Group Artifacts (N=4) | Final Individual Artifacts (N=9) |
|----------------------------|------------------------------|-------------------------------|----------------------------------|
| <b>Contextualization</b>   | No context                   | 1 (25.0%)                     | 0 (0.0%)                         |
|                            | Surface contextualization    | 3 (75.0%)                     | 1 (11.1%)                        |
|                            | Meaningful contextualization | 0 (0.0%)                      | 3 (33.3%)                        |
|                            | Context-driven task          | 0 (0.0%)                      | 5 (55.6%)                        |
| <b>Reasoning Potential</b> | Routine                      | 4 (100.0%)                    | 1 (11.1%)                        |

|                             |                            |           |           |
|-----------------------------|----------------------------|-----------|-----------|
|                             | Multi-step                 | 0 (0.0%)  | 4 (44.4%) |
|                             | Analytical / evaluative    | 0 (0.0%)  | 4 (44.4%) |
|                             | Generative / modeling      | 0 (0.0%)  | 0 (0.0%)  |
| <b>Context Authenticity</b> | Non-authentic / artificial | 1 (25.0%) | 1 (11.1%) |
|                             | Realistic yet generic      | 3 (75.0%) | 4 (44.4%) |
|                             | Realistic and relevant     | 0 (0.0%)  | 0 (0.0%)  |
|                             | Authentic and situated     | 0 (0.0%)  | 4 (44.4%) |

Table 2 documents clear structural variations between the spontaneous baseline tasks and the finalized individual products. In the contextualization dimension, context-driven problems rose to 55.6% in the individual stage, whereas bare mathematical tasks were entirely eliminated. In terms of reasoning potential, routine single-step formulations dropped from 100.0% to 11.1%, while multi-step problems (44.4%) and analytical/evaluative challenges (44.4%) emerged prominently. Generative modeling remained unrepresented (0.0%), indicating that teachers still found open-ended mathematical modeling challenging to construct. Authenticity distributions similarly revealed that authentic-situated problems increased to 44.4%, although realistic-generic tasks continued to constitute 44.4% of submissions.

These documented shifts highlight how structured *AI* interactions can stimulate teachers to formulate more complex instructional tasks. The transition away from surface-level story wrappers suggests that teachers engaged in more deliberate pedagogical thinking during the iterative revision cycle. However, these patterns should not be construed as a measured expansion of teacher pedagogical competence per se. Because the community service evaluation did not incorporate pre- and post-competency examinations or observational teaching rubrics, the recorded gains reflect task artifact quality within a supported setting. Descriptive artifact analysis confirms that the intervention environment enabled teachers to generate richer curricular materials than they produced spontaneously.

### Qualitative Evolution of Task Artifacts

To illustrate how task quality evolved across design stages, representative examples were examined to contrast initial baseline formulations with their subsequent collaborative and individual revisions. Baseline prompts routinely featured direct computational prompts, such as multiplying rows of cars ( $8 \times 5$ ) or calculating basic subtraction ( $25 - 9$ ). During collaborative *AI* prompting, teachers prompted the tool to generate Depok-specific settings, producing scenarios situated within local transit corridors and regional agro-tourism festivals. A concrete progression of how a routine car-parking prompt was adapted into situated local contexts is displayed in the following excerpt box.

| Soal Awal                                                                                                                     | Kelompok 1                  | Soal Berbantuan AI                                                                                                                                                                             |
|-------------------------------------------------------------------------------------------------------------------------------|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>Ditempat parkir terdapat 8 baris mobil. Setiap baris berisikan 5 mobil. Berapa jumlah mobil yang ada di tempat parkir?</p> | Versi 1                     | <p>Pada Festival Belimbing Dewa di Kota Depok, panitia menyusun 8 keranjang. Setiap keranjang berisi 5 buah belimbing. Berapa jumlah seluruh buah belimbing yang disiapkan panitia?</p>        |
|                                                                                                                               | Versi 2 (Masjid Kubah Emas) | <p>Saat kegiatan di kawasan Masjid Kubah Emas, terdapat 8 rak sandal. Setiap rak berisi 5 pasang sandal. Berapa jumlah seluruh pasang sandal yang ada di rak?</p>                              |
|                                                                                                                               | Versi 3 (Angkot Margonda)   | <p>Di sebuah terminal di Jalan Margonda, terdapat 8 angkot yang sedang menunggu penumpang. Setiap angkot membawa 5 penumpang. Berapa jumlah seluruh penumpang yang dibawa angkot tersebut?</p> |
| <b>Initial Problem:</b>                                                                                                       |                             |                                                                                                                                                                                                |

There are 8 rows of cars in a parking lot. Each row contains 5 cars. How many cars are there in the parking lot?

**AI-Assisted Problems:**

**Version 1 (Dewa Starfruit):**

At the Dewa Starfruit Festival in Depok City, the organizers prepared 8 baskets. Each basket contains 5 starfruits. How many starfruits did the organizers prepare in total?

**Version 2 (Golden Dome Mosque):**

During an activity in the area of the Golden Dome Mosque, there are 8 sandal racks. Each rack contains 5 pairs of sandals. How many pairs of sandals are there in total on the racks?

**Version 3 (Margonda Angkot):**

At a terminal on Margonda Street, there are 8 angkots waiting for passengers. Each angkot carries 5 passengers. How many passengers are being carried by the angkots in total?

**Figure 2. Collaborative AI-Assisted Contextual Transformation of a Baseline Problem**

Figure 2 showcases the immediate effect of *AI*-supported adaptation on narrative setting and local relevance. While the revised versions incorporated authentic Depok locations such as Margonda Street and regional fruit festivals, their underlying cognitive demand initially remained locked in routine single-step multiplication. This qualitative comparison illustrates a critical pedagogical distinction: embedding local nomenclature enhances surface relevance, but does not inherently elevate the cognitive reasoning demand of a mathematics task. Recognizing this limitation, participating teachers engaged in further dialogue to restructure subsequent prompts to target analytical thinking.

In later revisions, teachers successfully restructured the mathematical mechanics to require evaluative judgments and multi-step reasoning. For example, a baseline subtraction problem concerning animals was redesigned into an evaluative task requiring students to assess whether 12 families could adopt pets while retaining at least 15 animals at a community shelter. Another group expanded a stationery problem into a budgetary optimization challenge, asking pupils to determine the most cost-effective combination of school supplies within a fixed Rp100,000 allowance. This progression from simple computational queries toward justificatory prompts demonstrates that *AI* feedback served as a productive sounding board when guided by teacher intentionality (Hwang et al., 2024; Zhang et al., 2024).

**The Imperative of Human-in-the-Loop Verification**

Despite the communicative fluency of generative *AI*, workshop observations revealed that the tool frequently fabricated mathematically absurd or socially nonsensical scenarios. In one notable instance, the *AI* generated an adoption problem involving 250,000 rescue dogs at a local civic festival, asking whether 16 families could immediately adopt 850 dogs each. Structurally, the proposed task satisfied criteria for multi-step division and constraint analysis, yet its situational premises were completely detached from real-world plausibility. The complete text of this implausible *AI*-generated task artifact is presented in the following box.

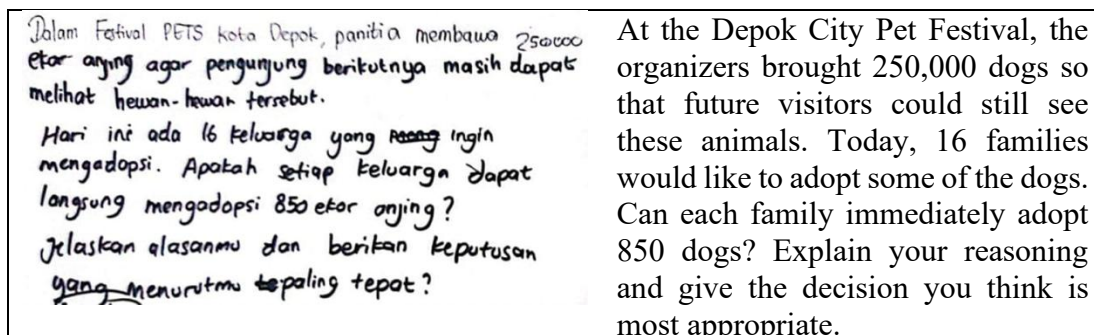


**Figure 3. Example of a Cognitively Demanding but Situational-Implausible *AI* Task Output**

Figure 3 demonstrates why algorithmic feedback must never be granted uncritical pedagogical authority in educational task design. When confronted with this task, the participating teachers did not accept the text uncritically; instead, they audited the numbers, recognized the physical absurdity of accommodating a quarter of a million animals, and rejected the prompt. This critical auditing episode exemplifies the human-in-the-loop principle advocated by international educational bodies (Miao & Cukurova, 2024; Miao & Holmes, 2023). Algorithmic tools generate permutations rapidly, but human teachers must supply the contextual gatekeeping, ethical appraisal, and mathematical sanity checks required for classroom-ready instructional tasks. Consequently, the teachers' rejection of the flawed task underscores that effective curriculum design depends on a systemic interplay between task structure, feedback design, and teacher judgment, rather than on the output of any single tool (He et al., 2026; Lee et al., 2024; Lucy et al., 2024; Modén et al., 2021; Rezat, 2026).

The qualitative findings indicate that generative *AI* facilitated task design through three complementary mechanisms: divergence, pedagogical contrast, and critical auditing. First, *AI* expanded the design space by providing multiple narrative scenarios from a single prompt, helping teachers move beyond conventional workbook archetypes. Second, *AI* generated contrasting variations that made differences in cognitive depth visible, enabling peer groups to compare routine calculations with evaluative tasks. Third, computational hallucinations and implausible outputs served as constructive learning triggers that compelled teachers to justify mathematical constraints and real-world boundaries. Consequently, the empowerment observed arose not from automated machine generation, but from teachers exercising collective professional judgment over algorithmic suggestions (Krushinskaia et al., 2025; Liu & Shang, 2025; Tan et al., 2025; Turvey & Pachler, 2025).

Several methodological and structural limitations must be acknowledged regarding these findings. The community service evaluation relied entirely on qualitative artifact analysis without independent double-coding, inter-rater reliability scores, or standardized scoring metrics. Furthermore, because the project did not include direct classroom trials with elementary school pupils, the empirical suitability and comprehensibility of these revised tasks for young learners remain unverified. Teacher participation was also confined to a single regional cohort in Depok City, which limits the transferability of the observed collaborative dynamics. Future community service and research initiatives should incorporate classroom implementation data and formal inter-rater reliability protocols to substantiate instructional impacts more robustly. Such longitudinal designs would also allow researchers to track whether the critical auditing practices observed here persist as teachers gain experience with generative tools.



## CONCLUSIONS

This community service program demonstrates that generative *AI* can function as an effective reflective feedback partner for elementary school teachers engaging in mathematical problem posing, provided it is embedded within a participatory, human-directed design framework. The intervention successfully assisted participating teachers in transforming routine computational prompts into more cognitively challenging and locally situated task artifacts. The documented artifact improvements highlight the viability of using *AI* to spark divergent instructional ideas, illuminate variations in cognitive demand, and stimulate collaborative critique. Importantly, the program also underscores that generative *AI* cannot replace professional teacher oversight, as human expertise remains essential for auditing mathematical soundness, numerical scale, and situational plausibility.

To ensure the sustainable adoption of these design practices, several realistic follow-up initiatives are recommended for school clusters and educational partners. Participating schools should establish collaborative lesson-design study groups where teachers curate shared task portfolios documenting original prompts, *AI* feedback dialogues, verification notes, and finalized instructional tasks. School leadership can support this process by introducing practical design checklists focusing on mathematical validity, real-world plausibility, and cognitive demand. Furthermore, priority should be given to conducting small-scale classroom trials of teacher-designed tasks to evaluate how elementary students interpret and resolve these situated mathematical problems. Future academic service programs should incorporate formal inter-rater reliability measures and longitudinal classroom observations to examine how sustained *AI*-assisted task design influences daily instructional practices.

## REFERENCES

- Brunetto, D., & Iacono, U. D. (2026). Modeling cycle and GenAI as resources for mathematics teachers' professional development. *Education Sciences*, 16(4), Article 504. <https://doi.org/10.3390/educsci16040504>
- Cai, J. (2022). What research says about teaching mathematics through problem posing. *Éducation et Didactique*, 16(3), 47–68. <https://doi.org/10.4000/educationdidactique.10642>
- Chavarría-Arroyo, G., & Albanese, V. (2023). Contextualized mathematical problems: Perspective of teachers about problem posing. *Education Sciences*, 13(1), Article 6. <https://doi.org/10.3390/educsci13010006>
- Chui, K. S. (2026a). *Using large language models as formative assessors evaluating students' mathematical modelling competence with ChatGPT* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.18961670>
- Chui, K. S. (2026b). *Using large language models as formative assessors evaluating students' mathematical modelling competence with ChatGPT* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.18961669>
- Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. <https://doi.org/10.54300/122.311>
- Glegola, A.-K., Jentsch, A., Ross, N., König, J., & Kaiser, G. (2026). The influence of teachers' dispositions on their use of mathematical problem-solving tasks: A Qualitative Comparative Analysis based on teachers' pedagogical content knowledge and



- beliefs. *Journal of Mathematics Teacher Education*.  
<https://doi.org/10.1007/s10857-026-09761-2>
- Hatisaru, V. (2020). Exploring evidence of mathematical tasks and representations in the drawings of middle school students. *International Electronic Journal of Mathematics Education*, 15(3), Article em0593.  
<https://doi.org/10.29333/iejme/8482>
- He, P., Li, Z., Wang, Z., Xiong, J., & Li, T. (2026). Judging the judges: Human validation of multi-LLM evaluation for high-quality K-12 science instructional materials. *arXiv*.  
<https://doi.org/10.48550/arxiv.2602.13243>
- Henderson, M., Bearman, M., Chung, J., Fawns, T., Buckingham Shum, S., Matthews, K. E., & de Mello Heredia, J. (2025). Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness. *Assessment & Evaluation in Higher Education*, 51(5), 863–878.  
<https://doi.org/10.1080/02602938.2025.2502582>
- Hwang, S., Xu, R., Yao, Y., & Cai, J. (2024). Learning to teach through problem posing: A teacher's journey in a networked teacher–researcher partnership. *The Journal of Mathematical Behavior*, 73, Article 101120.  
<https://doi.org/10.1016/j.jmathb.2023.101120>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kuhn, J. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274.  
<https://doi.org/10.1016/j.lindif.2023.102274>
- Krushinskaia, K., Elen, J., & Raes, A. (2025). Pre-service teachers' agency during their interactions with generative AI while designing for learning – a process view on Intelligent-TPACK. *Computers and Education Open*, 10, Article 100325.  
<https://doi.org/10.1016/j.caeo.2025.100325>
- Lee, J., Smith, D., Woodhead, S., & Lan, A. (2024). Math multiple choice question generation via human-large language model collaboration. *arXiv*.  
<https://doi.org/10.48550/arxiv.2405.00864>
- Lee, Y. (2024). The effects of functional moves in teacher questioning on students' contextualization of mathematical word problem solving. *Journal of Mathematics Teacher Education*, 28(1), 73–101. <https://doi.org/10.1007/s10857-023-09616-0>
- Liu, Y., & Shang, J. (2025). Charting diverse pathways to AI empowerment among teachers: A multi-case study in Chinese K-12 education. *International Journal of Chinese Education*, 14(3), 1–18. <https://doi.org/10.1177/2212585x251408661>
- Lucy, L., August, T., Wang, R. E., Soldaini, L., Allison, C., & Lo, K. (2024). Mathfish: Evaluating language model math reasoning via grounding in educational curricula. *arXiv*. <https://doi.org/10.48550/arxiv.2408.04226>
- Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000391104>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Modén, M. U., Tallvid, M., Lundin, J., & Lindström, B. (2021). Intelligent tutoring systems: Why teachers abandoned a technology aimed at automating teaching processes.



- Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 186–195. <https://doi.org/10.24251/hicss.2021.186>
- Parrish, C. W., & Bryd, K. O. (2022). Cognitively demanding tasks: Supporting students and teachers during engagement and implementation. *International Electronic Journal of Mathematics Education*, 17(1), Article em0668. <https://doi.org/10.29333/iejme/11475>
- Rezat, S. (2026). Students' mathematical feedback processes using a digital curriculum resource: A qualitative perspective. *ZDM – Mathematics Education*. <https://doi.org/10.1007/s11858-026-01821-4>
- Spreitzer, C., Straser, O., Zehetmeier, S., & Maaß, K. (2024). Mathematical modelling abilities of artificial intelligence tools: The case of ChatGPT. *Education Sciences*, 14(7), Article 698. <https://doi.org/10.3390/educsci14070698>
- Tan, S. C., Tan, Y. Y., Teo, C. L., & Yuan, G. (2025). Teachers' professional agency in learning with AI: A case study of a generative AI-based knowledge building learning companion for teachers. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.70013>
- Thomas, C. A., Berry, R. Q., & Sebastian, R. (2024). Examining the elements of culturally relevant pedagogy captured and missed in a measure of high-quality mathematics instruction. *ZDM – Mathematics Education*, 56(5), 953–964. <https://doi.org/10.1007/s11858-024-01595-7>
- Trevisani, F. de M., & Maltempi, V. (2026). Generative AI and pedagogical authorship in mathematics education. Dalam *WTM-Verlag eBooks* (hlm. 159–164). <https://doi.org/10.37626/ga9783959873680.0.27>
- Turvey, K., & Pachler, N. (2025). A topological exploration of convergence/divergence of human-mediated and algorithmically mediated pedagogy. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.70007>
- Zhang, L., Stylianides, G. J., & Stylianides, A. J. (2024). Enhancing mathematical problem posing competence: A meta-analysis of intervention studies. *International Journal of STEM Education*, 11(1), Article 48. <https://doi.org/10.1186/s40594-024-00507-1>