

Small vs. Large Language Models: Comparing Language Models in Generating Productive Failure Math Problems

Seyedahmad Rahimi, CLASS Lab, College of Lifetime Learning, Georgia Tech., srahimi@gatech.edu

Salah Esmaeiligoujar, Deniz Ercan, Maryam Babaei, Ran Gao
srahimi@ufl.edu, s.esmaeiligoujar@ufl.edu, deniz.ercan@ufl.edu, babaei.maryam@ufl.edu, gaoran@ufl.edu
School of Teaching and Learning, CLASS Lab, University of Florida

Abstract: Large language models (LLMs) such as GPT-4o can generate high-quality math problems but are resource-intensive and difficult to fine-tune. Small language models (SLMs) like Phi-3 may provide a more practical alternative, as they can run locally and be adapted for specific tasks. In this study, we generated 60 Productive Failure (PF) math problems, with half produced by GPT-4o and half by Phi-3. To ensure comparability, we controlled for grade level, Common Core standard, theme, interest, and prior knowledge. Trained raters evaluated the problems across five PF dimensions. Results showed that GPT-4o ($M = 13.38$; $SD = 1.88$) outperformed Phi-3 ($M = 10.67$; $SD = 2.73$), with a large effect size ($d = 1.16$). These findings indicate that although GPT-4o currently generates more effective PF problems, Phi-3 shows promising potential. With further fine-tuning and task-specific adaptation, SLMs like Phi-3 may offer a sustainable, cost-effective solution for educational problem generation.

Introduction

Large Language Models (LLMs) such as GPT-4o have become central to generating contextually rich and diverse math word problems (Naveed et al., 2023). However, LLMs' use comes with considerable financial and environmental costs (Morrison et al., 2025). From a global perspective, training LLMs consumes enormous computational resources, with datacenter-level operations driving energy demands and associated carbon emissions (Mehta, 2024). From a task-level perspective, for some specific purposes, using LLMs seems not appropriate. That is, we may use these very powerful models for tasks that could be done by much smaller models. Lastly, the use of LLMs, as of today, costs a considerable amount of money (particularly when text and image generation happen at the same time). These issues raise pressing concerns about the scalability of LLM-based AI-powered educational technology tools.

Small language models (SLMs), in contrast, offer a promising alternative. Defined broadly as models under 10 billion parameters that can operate on consumer-grade hardware, SLMs are increasingly capable of handling specialized tasks with competitive accuracy and efficiency (Belcak et al., 2025). SLMs are less costly to run, can be finetuned with relatively modest computational resources (Dettmers et al., 2023), and may be deployed locally, which can enhance both privacy and sustainability. Their task-specific focus also aligns well with many educational applications, which often require a narrower scope and curricular alignment rather than general conversational ability (Lu et al., 2024).

The current study investigates whether SLMs can rival LLMs in producing high-quality math problems designed for teaching through Productive Failure (PF; Kapur, 2014; Kapur & Bielaczyc, 2012). PF is a teaching method that intentionally confronts learners with challenging problems designed for failure before formal instruction. These problems leverage failure or struggle to activate prior knowledge, encourage exploration of multiple solution pathways, and foster deeper conceptual understanding (Kapur, 2014; Kapur & Bielaczyc, 2012). Research has shown that PF supports robust learning outcomes across domains (Loibl et al., 2017) and that PF usually outperforms direct instructions (Sinha & Kapur, 2021; Darabi et al., 2018). The word problems suitable for PF are usually long, story-like, and effectively engaging. Creating these problems for teachers is not an easy task. Therefore, through detailed prompt engineering, we used LLMs and SLMs to help teachers.

To operationalize PF problems' quality, we employed a rubric based on the literature of PF which captures five critical dimensions: activation of prior knowledge (APK; i.e., a PF problem should include familiar concepts, sweet spot calibration (SSC; a PF problem should include concepts that learners have mastered before so the problem is approachable), affective engagement (AE; a story with theme and information that aligns with students' interests can engage learners affectively), multiple representations (MR; a PF problem should encourage a wide range of representations and solution methods), and open-endedness (OE; a PF problem should be open-ended and ill-structured to allow for creative exploration and non-canonical ideas). Each of these facets could be scored for each problem as 1, 2, or 3, making the minimum score as 5 and the maximum score as 15. This rubric was applied to evaluate problems generated by GPT-4o and Phi-3. Phi-3 (mini; 3.8B parameters), a

fast, efficient model trained on high-quality synthetic data and optimized for real-time, on-device use. The central research question for this study is: *Is there any difference between GPT-4o and Phi-3 in the quality of PF problems that they generate? If so, how large is this effect?*

Method

Problem generation

We used the two models' API to generate 60 algebra problems (GPT-4o = 30; generated by Phi-3 = 30). Both models received identical prompts and problem contexts. To ensure fairness across models, both models received the same structured prompt with six components: (1) *roles and objectives*, defined the model's task as an educational design assistant generating PF math problems; (2) *response rules*, specified the narrative format, tone, open-endedness, and restrictions; (3) *PF overview*, described PF's phases and purpose; (4) *a detailed rubric*, outlined six evaluation criteria for evaluating the quality of PF problems, each scored on a three-point scale; (5) *human-generated examples*, illustrated five example math problems to show the expected quality and the structure of PF problems; (6) *instructions for generating new problems*, guided the model to create a new, open-ended narrative problem aligned with a selected Common Core Standard, grade level, theme and interest, and explicitly listed prior knowledge concepts.

In generating the 60 math problems, we controlled for confounding variables by holding constant grade level, domain, Common Core standard, thematic context, student interests, and prior knowledge concepts. We used topics from 7th grade ratios and proportional relationships (CCSS.Math.Content.7.RP.A.2), with prior knowledge drawn from unit rates (7.RP.A.1) and operations with rational numbers (7.NS.A.1). The thematic context was sports, specifically basketball and scenarios involving Stephen Curry. Each problem was assigned to a unique identifier and stored in a blinded database for evaluation (i.e., raters did not know the model with which the problem at hand was generated).

Evaluation process

Evaluation was conducted by at least two trained human raters. Each problem was scored independently by two raters using the PF rubric (see Table 1). Discrepancies of ≥ 5 points triggered adjudication by a third rater (16 cases), with the final score taken as the mean of the two closest ratings. This process ensured reliability and minimized outlier effects. Data were analyzed with independent-samples t-tests comparing GPT-4o and Phi-3 on overall and subscale scores performance, with effect sizes (Cohen's *d*) and 95% confidence intervals reported.

Table 1
The Rubric for Assessing Productive Failure Problems

Dimension	1 (Low)	2 (Moderate)	3 (High)
Activation of Prior Knowledge	No prerequisite concepts	Includes one prerequisite concept	Includes two or more prerequisite concepts
Sweet Spot Calibration	Too easy or too difficult: disengagement	Moderate challenge with some failures	Well-calibrated: challenging but manageable, leading to productive engagement
Affective Engagement	Not relatable or meaningful: fails to engage	Somewhat engaging with relatable/realistic elements	Highly engaging, presented in meaningful narrative or real-world scenario
Multiple Representations & Solutions	Few or no representations or solution methods	Some diversity but limited	Wide range of representations and solution methods; fosters creativity & collaboration
Open-endedness / Ill-structuredness	Simple, well-structured, single path	Moderate complexity with limited variables	Highly complex, requiring integration of multiple variables and representations

Results

Table 2 summarizes the means, standard deviations, and t-test results for GPT-4o and Phi-3 across all rubric dimensions. The Final Score was significantly higher for GPT-4o ($M = 13.38$; $SD = 1.88$) than for Phi-3 ($M = 10.67$; $SD = 2.73$), with a large effect size ($d = 1.16$), which shows that GPT-4o demonstrated stronger overall performance across all rubric dimensions. The largest effects were observed for multiple representations ($d = 1.65$) and open-endedness ($d = 1.58$), suggesting that GPT-4o is particularly adept at generating problems that integrate diverse representational forms and allow for multiple solution pathways. Differences in affective engagement did not reach significance ($p = .06$), which indicates comparable performance between the two models on this dimension.

Table 2
T-test Results for GPT-4o vs. Phi-3 on PF Rubric Dimensions

Measure	GPT-4o M (SD)	Phi-3 M (SD)	t	p	d	95% CI
APK	2.65 (0.35)	2.42 (0.53)	2.02	0.049	0.52	[0.00, 0.47]
SSC	2.52 (0.55)	2.08 (0.62)	2.87	0.01	0.74	[0.13, 0.74]
AE	2.75 (0.47)	2.45 (0.71)	1.93	0.06	0.50	[-0.01, 0.61]
MR	2.70 (0.43)	1.83 (0.61)	6.40	<0.001	1.65	[0.59, 1.14]
OE	2.77 (0.47)	1.88 (0.64)	6.11	<0.001	1.58	[0.59, 1.17]
Final S	13.38 (1.88)	10.67 (2.73)	4.49	<0.001	1.16	[1.50, 3.93]

Note. APK = Activation of Prior Knowledge; SSC = Sweet Spot Calibration; AE = Affective Engagement; MR = Multiple Representations; OE = Open-endedness. Final = total score across all rubric dimensions. M = mean; SD = standard deviation; d = Cohen's d effect size; CI = confidence interval; Final S = Final Score; Significant results are shown in bold.

Figure 1 shows the distribution of Final Scores for GPT-4o and Phi-3. The violin plot reveals the overall distribution, while the box plot shows medians and quartiles. GPT-4o's scores cluster tightly near the higher end (around 13–15), with less variability, whereas Phi-3's scores spread more widely, ranging from very low (around 6) to high values near 11–14. These results indicate that GPT-4o performed more consistently at a higher level compared to Phi-3, which is more variable, with some very weak outputs and some strong ones.

Figure 1
Box/violin plot of Final Scores by Model

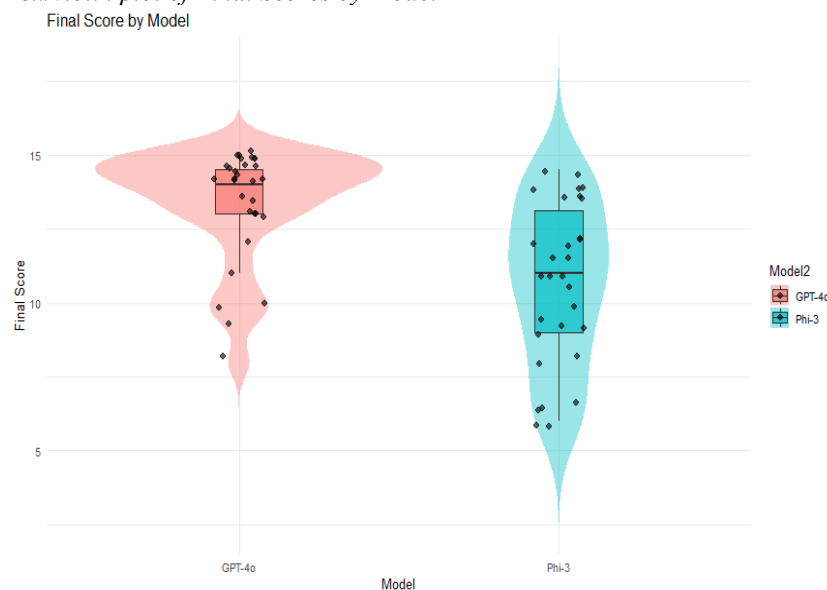
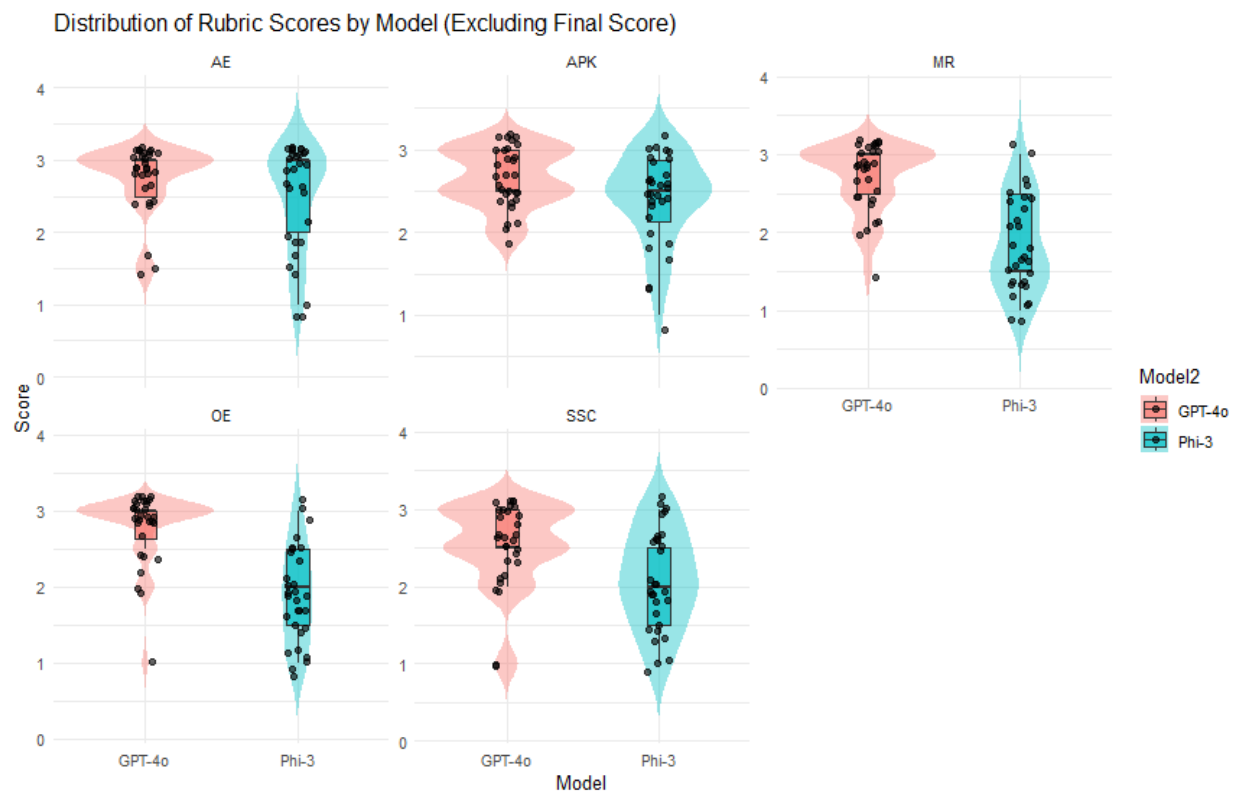


Figure 2 compares GPT-4o and Phi-3 across all rubric dimensions. In nearly all categories, GPT-4o shows higher medians and tighter distributions that suggest greater consistency. Phi-3 shows wider distributions with lower medians. This result indicates less stable performance for Phi-3. Notably, GPT-4o's advantage is most visible in MR and OE. These findings highlight areas where GPT-4o delivers stronger and more reliable problem quality.

Figure 2
Box/violin plots of Subscores by Model (APK, SSC, AE, MR, OE)



Discussion and conclusion

In this study, we investigated the difference in generating 7th grade PF math problems using GPT-4o (an LLM) and Phi-3 (an SLM). Our findings show that GPT-4o generates PF problems with higher quality than Phi-3 overall and across most dimensions, with the largest gaps in multiple representations and open-endedness—key to PF's focus on exploration and non-linear problem-solving (Kapur, 2014). These findings align with prior work that showed LLMs' strength in integrating complex representational structures due to diverse multimodal training (Naveed et al., 2023). While Phi-3 performed comparably on affective engagement and only slightly lower on activation of prior knowledge, GPT-4o still produced more pedagogically sophisticated tasks. Yet Phi-3's efficiency and smaller size highlight the promise of SLMs as sustainable alternatives.

Future studies should look into enhancing SLMs' performance by improving the prompt engineering and integrating targeted fine-tuning or retrieval-augmented generation (Borgeaud et al., 2022; Belcak et al., 2025; Shone, 2024). Moreover, instead of using only one SLM, future studies could engineer a multi-model system with multiple SLMs that support and communicate with each other for optimal performance.

The study is limited to sixty problems within one domain and grade level, suggesting the need for broader evaluations, classroom-based teacher reviews, and refinement of SLM prompts. Overall, GPT-4o currently outperforms Phi-3, especially in representational reasoning, but SLMs' efficiency underscores their potential for accessible and ecologically responsible educational applications.

References

- Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., ... Lin, Y. C. (2025). *Small language models are the future of agentic AI*. arXiv preprint arXiv:2506.02153
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... Rae, J. (2022). Improving language models by retrieving from trillions of tokens. *Advances in Neural Information Processing Systems*, 35, 22023–22037.
- Brookhart, S. M. (2018). *How to create and use rubrics for formative assessment and grading* (2nd ed.). ASCD.
- Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, 66(5), 1101–1118.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088–10115.
- Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, 38(5), 1008–1022.
- Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, 21(1), 45–83.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863.
- Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, 29(4), 693–715.
- Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., ... Xu, M. (2024). *Small language models: Survey, measurements, and insights*. arXiv preprint arXiv:2409.15790.
- Mehta, S. (2024, July 4). *How much energy do LLMs consume? unveiling the power behind ai*. ADaSci. <https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/>
- Morrison, J., Na, C., Dettmers, T., Strubell, E., Fernandez, J., & Dodge, J. (2025). *Holistically evaluating the environmental impact of creating language models*. arXiv preprint: 2506.02153.
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... Mian, A. (2023). *A comprehensive overview of large language models*. arXiv preprint arXiv:2307.06435.
- Shone, O. (2025, October 6). *Explore AI models: Key differences between small language models and large language models*. The Microsoft Cloud Blog. <https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/11/11/explore-ai-models-key-differences-between-small-language-models-and-large-language-models/>
- Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, 91(5), 761–798.