EPFL

CHILI
computer-human interaction
in learning and instruction

LEARN
Center for Learning Sciences

Exploring LLMs in Storytelling and Q&A Generation for Education Using a Social Robot

Author:
YASMINE CHAKER

Supervisor(s):
Daniel Carnieto Tozadore

Professor:
Pierre Dillenbourg

January 10, 2025

Contents

1 Introduction

Large Language Models (LLMs) are revolutionizing Elementary Education by enhancing student engagement and skill development, especially when paired with social robots. LLMs aid in improving reading, writing, and critical thinking by offering tailored feedback and generating texts that foster deeper understanding [6]. Additionally, social robots contribute by promoting problem-solving, creativity, and collaboration, while boosting student motivation and computational thinking [4]. The challenge remains: how can these technologies be seamlessly integrated into education in schools to maintain engagement and ensure effective knowledge transfer?

One promising approach to integrating LLMs and social robots in education is through storytelling. Storytelling has been shown to enhance children's creativity, engagement, and learning outcomes. For example, children interacting with an LLM-driven storytelling robot demonstrated increased creativity, particularly in fluency, flexibility, and elaboration [9]. Additionally, storytelling fosters a meaningful, safe, and resource-rich environment, promoting deeper engagement and a stronger sense of community in the classroom [1].

Numerous scientific studies demonstrate how storytelling, enhanced by LLMs and social robots, can effectively support children's education. Storypark leverages an interactive storytelling approach, where children actively contribute through guided questioning and sketch-based interactions, leading to improved engagement and comprehension of story concepts [10]. Similarly, Mathemyths integrates mathematical language into narratives through co-creative storytelling, demonstrating that prompt engineering can optimize LLMs for educational outcomes comparable to human-guided storytelling [11]. StoryBuddy highlights the value of human-AI collaboration, offering flexible parental involvement and automatic question generation to tailor educational goals and track progress [3]. Lastly, a Question-Answer Pair Generation system presented in "It is AI's Turn to Ask Humans a Question" enhances reading comprehension by generating targeted QA pairs from children's storybooks, emphasizing the importance of education-specific models and datasets for better pedagogical performance [2].

In this project, we choose a version of Phi-3-Mini[7] which was fine-tuned on an educational dataset and proven to perform as a robust educational tutor. We utilize an interactive interface previously developed in our laboratory. The interface is originally powered by OpenAI's GPT-3.5-turbo API (referred to as "Turbo"). Our objective is to firstly integrate our fine-tuned model (referred to as "Phi") and then to test and compare the story and question generation capabilities, based on the course content, of three models: Phi, Turbo, and teacher-generated content (referred to as "Human"). These models are evaluated in a real classroom setting with three groups of year-4 children aged 8-9, using UBTECH's Alpha-Mini Humanoid Robot for narration. We aim to answer the following research questions:

- **RQ1:** Can LLMs, with appropriate prompting, match human performance in clarity and adaptation of stories and questions to children's comprehension levels?
- **RQ2:** Do children have a preference for one model over the others in terms of enjoyment?
- **RQ3:** Is the fine-tuned model perceived as more educational by the children compared to the other models?

2 System Design

2.1 Model Selection

As the number of large language models continues to grow, we need to define the criteria for selecting the model best suited to our project. It should be open-source to allow access to its weights, available for research use to enable practical applications, lightweight to fit in memory and compute-constrained environments, and capable of handling a large context length to accommodate long prompts when necessary. These criteria narrow down the available options, and we initially select Phi-3-Mini-128K-Instruct[7], a 3.8 billion-parameter, lightweight, state-of-the-art open model trained on the Phi-3 datasets. Building on previous work conducted in the lab, we have evidence that this model, when fine-tuned on an educational dataset, performs better on tasks requiring the integration of detailed, topic-specific knowledge and logical reasoning. Due to time and computational constraints, as well as the fact that the fine-tuned model perfectly aligns with the goals of our project, we opt for it.

Technical Details of Fine-Tuning
A QLORA[5] approach was used to fine-tune the Phi-3-Mini-128K-Instruct model on an educational dataset. Specifically, it was fine-tuned on a sample from Cosmopedia[8], a dataset of synthetic textbooks, blog posts, stories, WikiHow articles, and posts generated by the Mixtral-8x7-Instruct model. Samples from Stanford courses, Khan Academy, WikiHow, and other educational sources were used to prompt the model and generate a dataset covering 145 different topics, aimed at various audiences ranging from elementary school students to researchers. A subsample of the dataset, containing 2,138 data points, was randomly selected, and the QLORA adapters were fine-tuned on this subsample.

2.2 Integration and Modifications of the Existing Interface

2.2.1 Overview of the Existing Interface

The existing interface was developed in the context of a semester project in the lab and is intended to be used with any social robot. The interaction between the teacher and the robot is facilitated through a web user interface designed to manage AI-driven story generation and Q&A sessions. The interface, powered by OpenAI's GPT-3.5-turbo API,

enables teachers to customize the AI model's behavior, including language preferences and generation options, and provides varying levels of AI assistance.

The story generation process includes multiple AI assistance levels, allowing the teacher to choose how much support they need, from creating complete stories independently to offering prompts for AI-generated content. The interface also supports age group customization, tailoring story complexity and themes to different developmental stages, such as toddlers, preschoolers, early elementary, late elementary and preteens.

Post-story generation, teachers can refine stories through additional AI-assisted tuning options. Following the storytelling, the interface supports a Q&A session, where teachers can either input their own questions and answers or rely on AI to generate them.

2.2.2 Interface Adaptation for Story and Question Evaluation

As the goal of our project is to assess the quality of the generated stories and questions of different models rather than to validate the user experience with the interface, certain adjustments were made to align with this purpose. Specifically, instead of allowing the choice of various parameters, we set them to default values as follows:

- The language is set to English.
- The AI assistance level is set to 4, the highest level, meaning that the user only provides the topic for the story to be generated by the model.
- The age group is set to Late Elementary, as the experiments were conducted in an elementary school. Younger groups are too young to provide accurate judgments on how they perceive the stories and questions, whereas Late Elementary students are better suited for this task.

Integration of the Phi Model
We extract the relevant parts of the code from the existing interface discussed in the previous subsection. This includes methods that define the characteristics of the target age groups, the characteristics of the generated stories and questions tailored for those groups, and the methods that generate stories based on a given topic and questions based on the story.

Story characteristics for the Late Elementary age group: Middle-grade books with more complex narratives, themes of friendship, adventure, and personal growth, as well as exploration of emotions and relationships.

We first load the fine-tuned model defined in 2.1. We modify the methods so that our model generates the outputs, setting the word count to 500 and the temperature to 0.5 for the stories. The temperature parameter controls the balance between predictability (when

set to 0) and creativity (when set to 2). Most of the prompts used remained consistent with those used with the GPT API, with one key adjustment: instead of including specific examples of stories, questions, and answers in the prompt, we kept only the structure of the examples without the actual text. This change was necessary because the Phi model struggles to distinguish between the example text and the generated content, unlike GPT. Despite this, the model still performed well, as it was able to understand both the structure and the task effectively, resulting in no deterioration of the output quality.
The code can be found in this link to Github repository.

3 Experiments

3.1 Preparations and Material Generation

After contacting the International School and explaining our project, we received approval from the institution, and they obtained consent from the parents of the participating students. The teachers provided us with detailed files outlining the topics and objectives for each day of the week. We asked them to generate stories and questions based on the provided material. The detailed instructions given to the teachers can be found in Appendix A.1.

- *Topic 1 (Monday):* Food chains—identifying producers, predators, and prey.
- *Topic 2 (Tuesday):* Environmental changes caused by fire.
- *Topic 3 (Thursday):* The importance of working together to reduce environmental impacts.

On our side, we generated stories and corresponding Q&As for each topic using the GPT API and our fine-tuned Phi model. These were prepared in advance to ensure seamless progress of the activities without delays or interruptions.

3.2 On-Site Experimentation and Procedure

The experiments were conducted in a classroom of the International School over one week, with a break day in the middle. In Figure 1, the detailed session repartition is shown. Each session lasted approximately 50 minutes.

- **Age group of participants:** 8–9-year-olds (Year 4 in Elementary School).
- **Number of participants:** 45 students, of which 7 missed one or more sessions. Their data was excluded from the analysis, leaving a final sample of 38 students: 14 in Group 1, 14 in Group 2, and 10 in Group 3.

The three models used were:

- **Phi:** The fine-tuned model defined in section 2.1.
- **Human:** Content generated by a teacher.
- **Turbo:** Stories and Q&As generated using the GPT API.

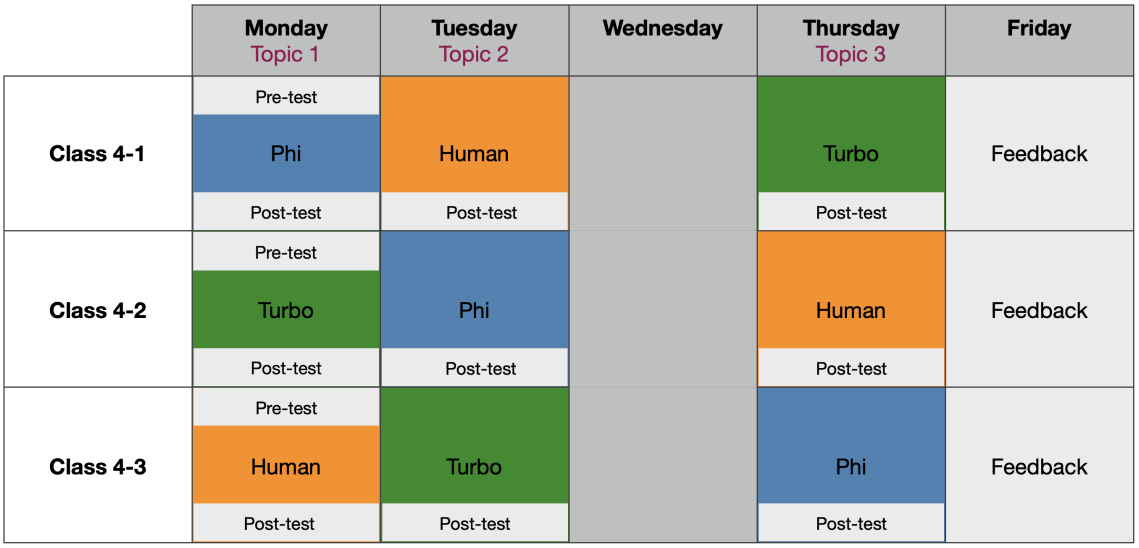


Figure 1: Session Repartition Across Groups and Days

Pre-Test: Filled on the first day at the start of the session. The pre-test survey evaluated children's familiarity with social robots, their attitudes toward them, and their preferences for interacting with robots in various settings. It contained 11 questions, ranging from prior experiences with robots to preferences for educational and social contexts.

Session Flow: The Alpha-mini robot greeted the students and introduced itself as a visitor from another galaxy. It asked if it could join their class for a week and expressed enthusiasm about learning from them about friendship and collaboration. Each day, the children were unaware of which model was generating the stories.
The daily session proceeded as follows:

- The robot narrated the story for the day's topic once.
- It asked three questions, allowing time for students to respond by writing down answers on the notebooks we distributed.
- The robot repeated the story and the questions, allowing students to review their initial responses and then provided the correct answers.

Post-Test (Daily Appreciation Survey): At the end of each session, students completed a survey with four questions:

1. How much did you like this activity?
2. How much do you think you learned with this activity?
3. How much did you like the robot's questions and answers?
4. How much did you prefer this activity compared to the one done yesterday? (not asked on Monday).

Responses were recorded on a 1–10 scale.

Final Feedback (Global Appreciation Survey): On the last day, the students listened to the stories generated by each model again and rated their enjoyment on a scale of 1–5 for each story. Afterward, they ranked the models from 1 to 3 based on their overall preference, with 1 being the most preferred. After finishing the survey, we introduced the three models to the students.

4 Results and Discussion

In order to perform analysis on the collected data, we began by digitizing the surveys filled out by the students. We also graded their answers (to the three questions) that they wrote in the notebooks by comparing them with the correct answers generated by the models. When digitizing the notebooks, we only reported the number of correct answers of each student in each session. Next, preprocessing was carried out to exclude the data of students who were not present in all four sessions, and to replace the few missing values in the remaining data with the mean of the respective column.

To report statistical significance, non-parametric tests were conducted due to the non-normality of the data distribution. The Kruskal-Wallis test was used to assess the differences between the three models, followed by Dunn's test for pairwise comparisons. Significant differences are indicated by stars in the plots, where '*' denotes $p < 0.05$, '**' denotes $p < 0.01$, and '***' denotes $p < 0.001$ (p being the p-value).

4.1 Pre-test Analysis

The eleven questions included in the pre-test survey are provided in Appendix A.2. In this section, we analyze the results of six specific questions:

Q1: Have you ever interacted with a social robot?

1. Never
2. Once or twice
3. More than three times in my life
4. Every week

Q2: How much do you like social robots?

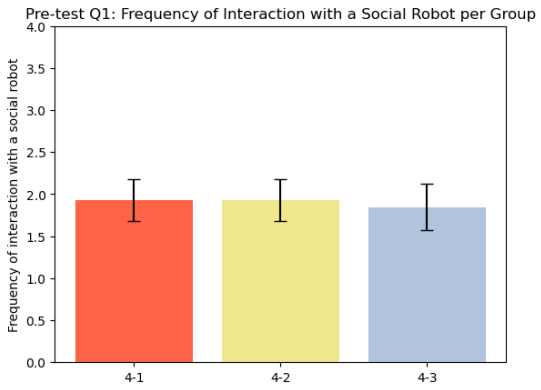
How much would you like to have a social robot assist you in the following activities:
Q3: Helping with your tasks/lessons in the classroom?
Q4: Helping with your lessons at home (homework)?
Q8: Which of these lessons do you like the most?
Q9: Which of these lessons do you think a robot can help you learn the most?
Mathematics & Technologies — Languages & Literature — Social Sciences & Arts

For Questions 2, 3, and 4, responses were recorded on a 1-5 scale. For Questions 8 and 9, students were allowed to select more than one subject.

We conducted an analysis to determine whether the three groups were comparable in terms of their familiarity with social robots, preferences, and their willingness to have a social robot assist them with schoolwork. Additionally, we examined the subjects students liked the most and subjects they believed a social robot could help them learn the most.

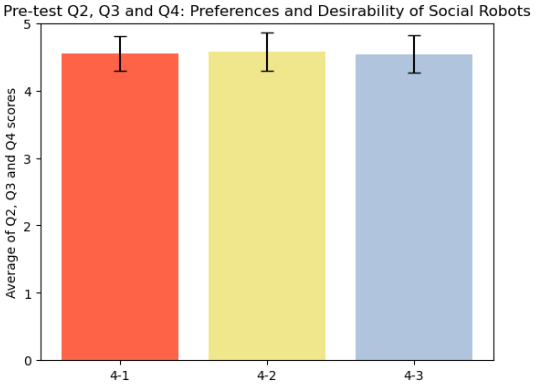
In Figure 2a, the three groups show comparable averages for the number of times they had interacted with social robots, with the average being "once or twice." In Figure 2, it is evident that all three groups rate their enthusiasm for social robots and their willingness to have one assist in their education highly, with an average score of 4.5 out of 5. This indicates that the majority were highly enthusiastic about integrating a social robot into their learning process.

In Figure 3, Mathematics & Technologies emerges as the most liked subject across all three groups, either alone or in combination with Social Sciences. Similarly, in Figure 4, Mathematics & Technologies is perceived as the subject where a robot could provide the most help. For Group 1, it is equally positioned with Languages & Literature.



(a)

1: Never, 2: Once or twice, 3: More than 3 times in my life, 4: Every week



(b)

Figure 2: Familiarity with Social Robots (a) and Preferences & Desirability of Social Robots for Education (b) per Group

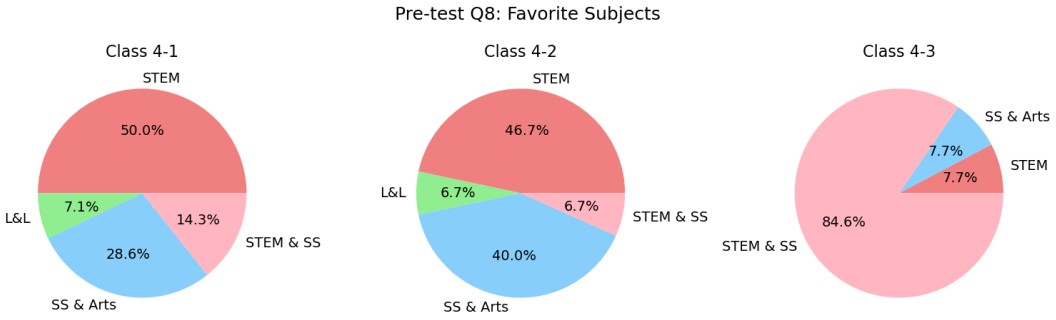


Figure 3: Students' Favorite Subjects
STEM: Mathematics & Technologies, L&L: Languages & Literature, SS: Social Sciences

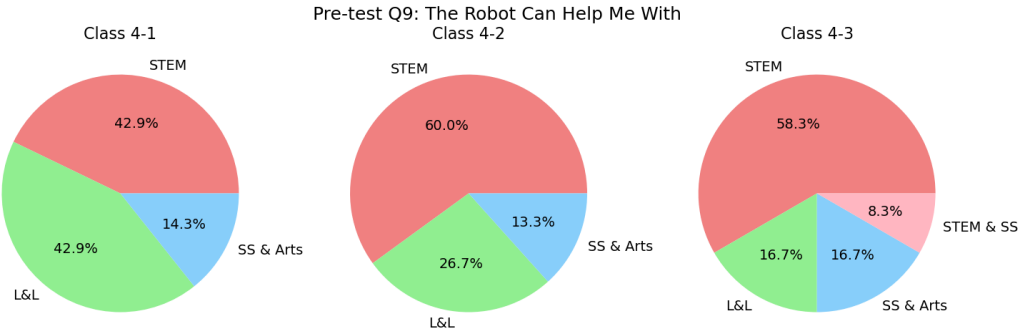


Figure 4: Subjects that the Students Think the Robot Can Help them Learn the Most
STEM: Mathematics & Technologies, L&L: Languages & Literature, SS: Social Sciences

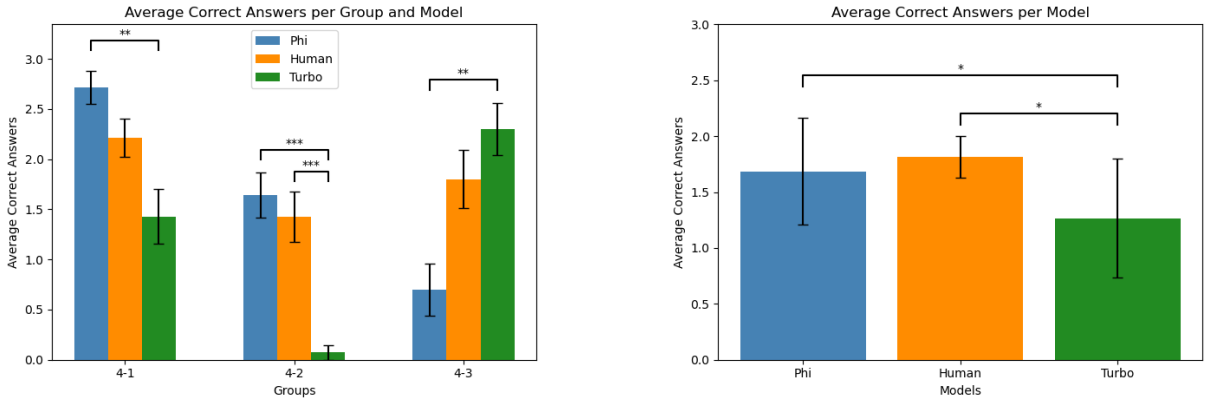
4.2 Average Correct Answers

In each session, students were asked three questions related to the story narrated by the robot. We compared the average number of correct answers obtained by the students in each session and each group. This metric allows us to evaluate how clear and well-adapted the stories and questions generated by each model were to the children's comprehension level.

In Figure 5a, we observe that the same pattern is repeated in the first two groups, where there are more correct answers with Phi than with Human, and more with Human than with Turbo. In contrast, group 3 shows the completely opposite pattern. There are two extreme values worth explaining. The first is the very low score for Turbo in group 2, where the generated text used difficult terms that were not suited for year 4 children. The second is the low score for Phi in group 3, where the generated questions were not straightforward and difficult to answer based on the story. An interesting point to note is the relatively small variance in the average number of correct answers with the stories generated by the teacher (Human model). This is due to the lack of randomness in the generation process, resulting in all stories and questions having a consistent style and level

of clarity. This is not the case for texts generated by LLMs.

In Figure 5b, we observe a statistically significant difference between Phi and Turbo, as well as between Human and Turbo, but not between Phi and Human. This indicates that Phi's generations can be comparable to Human generations in terms of clarity and adaptation to the specific comprehension level of the children.



(a) (b)

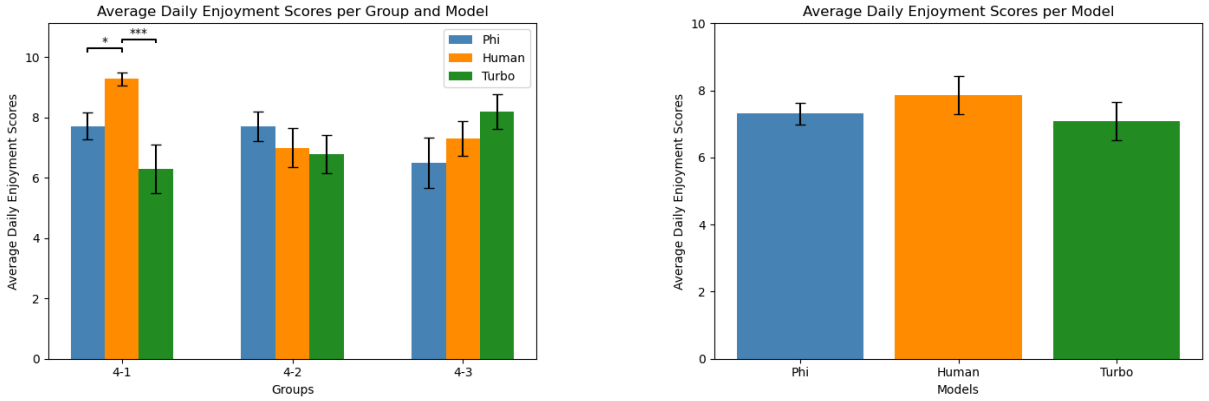
Figure 5: Average Correct Answers per Group (a) and per Model (b)

4.3 Enjoyment Scores

At the end of each session, the students were asked to rate how much they liked the activity on a 1-10 scale. This measured how enjoyable the stories and questions generated by each model were, as well as how enjoyable it was to have a robot narrate them.

In Figure 6a, we observe that the average values are in the upper half of the scale. The patterns for groups 2 and 3 are similar to those in the previous plot, although there are no statistically significant differences between the three models in this plot. For group 1, the students rated the Human model as more enjoyable than the Phi model, although they performed less well with it. This was because the story contained their names (many of them noted this reason when asked to rank the models on the last day, and they also seemed very excited when listening to the story).

In Figure 6b, we observe that the average enjoyment score is slightly higher for the Human model than for the Phi model, and slightly higher for the Phi model than for the Turbo model. However, these differences are not statistically significant. It is safe to conclude that the children enjoyed having the robot narrate stories for them with the three models.

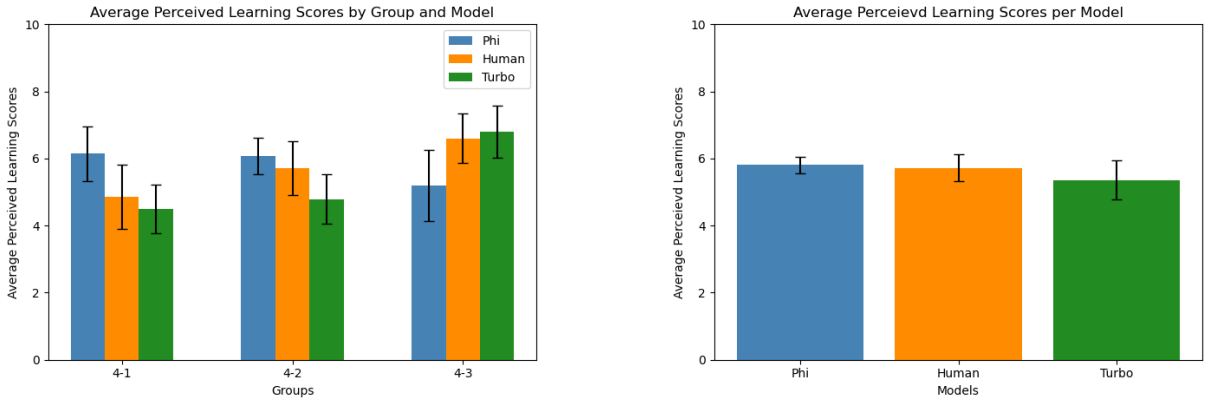


(a) (b)

Figure 6: Average Daily Enjoyment Scores per Group (a) and per Model (b)

4.4 Self Reported Learning Perception

At the end of each session, the students were asked to rate how much they felt they learned from the activity on a 1-10 scale. This rating reflects how useful they thought the session was in helping them learn new things.



(a) (b)

Figure 7: Average Perceived Learning Scores per Group (a) and per Model (b)

In Figure 7a, patterns similar to those in Figure 5a are observed across the three groups. However, the differences between the models are not statistically significant. This suggests that students may have felt they learned more when they understood the story better and were able to answer more questions correctly. To explore this further, we computed the Spearman correlation coefficients between the number of correct answers and the perceived learning scores for each model and group. In Table 1, the correlation values and their corresponding p-values are shown. Notably, the only statistically significant positive correlation is with the Phi model for Group 2, with a correlation coefficient

of 0.57.

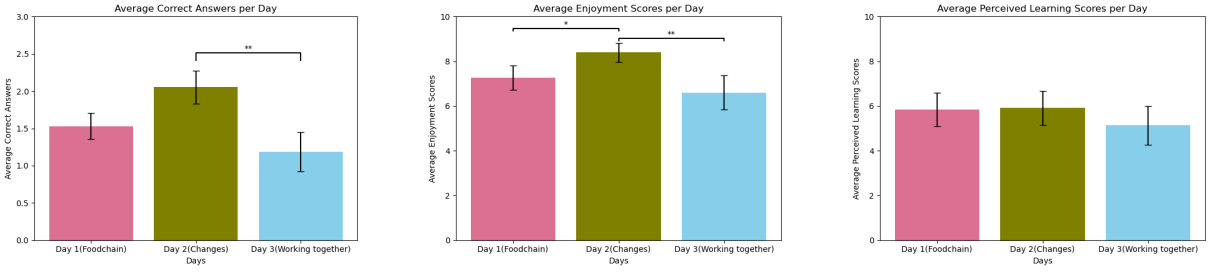
In Figure 7b, the average perceived learning score is slightly higher for the Phi model compared to the Human model and slightly higher for the Human model compared to the Turbo model. However, these differences are also not statistically significant. The averages are close, ranging between 5 and 6, suggesting that the activity may not have been perceived as highly effective in teaching the children new lessons.

Table 1: Correlation Coefficients (r) and p-values (p) Between Correct Answers and Perceived Learning Scores for Each Model Across Groups

	Phi	**Human**	**Turbo**
4-1	$r = -0.04, p = 0.891$	$r = -0.06, p = 0.842$	$r = -0.17, p = 0.567$
4-2	$\mathbf{r = 0.57, p = 0.032}$	$r = -0.01, p = 0.981$	$r = 0.07, p = 0.812$
4-3	$r = 0.24, p = 0.500$	$r = -0.52, p = 0.122$	$r = -0.32, p = 0.375$

4.5 Analysis by Day/Topic

Each day corresponds to a different topic, making the analysis per day equivalent to the analysis per topic. We aim to determine whether the days influence students' performance—specifically, whether their performance improves over time as they become accustomed to the robot and its storytelling method. Additionally, we want to assess whether enjoyment decreases over time, suggesting that the novelty of the robot fades, reducing its appeal and excitement.



(a) (b) (c)

Figure 8: Average Correct Answers (a), Enjoyment Scores (b) and Perceived Learning Scores (c) per Day/Topic

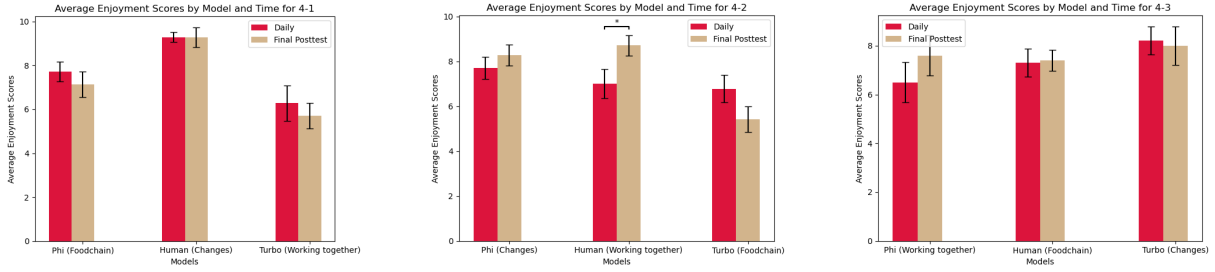
In Figures 8a and 8b, we observe that the second day, which focused on the topic of environmental changes, yielded the highest scores in both correct answers and enjoyment. These differences are statistically significant, indicating that the scores were influenced by the generated stories and questions -whose quality depended on the topics, rather than the order of the days. This suggests that the stories and questions for the environmental changes topic were well-suited to the children's comprehension level and were the most

enjoyable for them. Conversely, for the topic of "Working Together", the models failed to produce clear and engaging stories and questions. Notably, Phi achieved its lowest scores with group 3 during the "Working Together" topic (5a,6a,7a). This could be due to the non-scientific nature of the topic, making it more challenging to teach effectively.

In Figure 8c, we observe no significant differences in the average perceived learning scores across the days. However, day 2 shows a slightly higher average than the other two days, aligning with our previous observations.

4.6 Consistency between Daily and Final Enjoyment Scores

In the final feedback, students rated the stories generated by each model. We compared the scores from the daily feedback, which reflected their appreciation of the activity, with the scores from the final feedback to check for consistency. Pairwise analysis using the Mann-Whitney U test revealed a single statistically significant difference: in group 2 (Figure 9b), the story created by the teacher received a higher enjoyment score in the final feedback compared to its daily score. This suggests that the students might have enjoyed the story more without the accompanying questions. Besides, the story created by the teacher included the name of their school, which may have contributed to their increased interest.



(a) (b) (c)

Figure 9: Comparison between Daily and Final Average Enjoyment Scores for Group 1 (a), Group 2 (b) and Group 3 (c)

4.7 Students' Appreciation of the Models

On the last day, after listening to the stories again, students rated them and ranked the three models. As shown in Figure 10, the Human model was ranked first by the majority, with 21 students selecting it as their top choice. This was followed by the Phi model, chosen by 12 students, while the Turbo model ranked last, with only 5 students ranking it as the best.

These results suggest several factors influencing students' preferences. First, the teacher's story was highly personalized, incorporating the school's name and references to students

and their friends, making it more relatable and engaging. Second, the teacher's understanding of the students' abilities allowed her to tailor the story to match their current comprehension level, further increasing its appeal. Personalization, context, and familiarity appear to play key roles in shaping students' engagement and preferences.

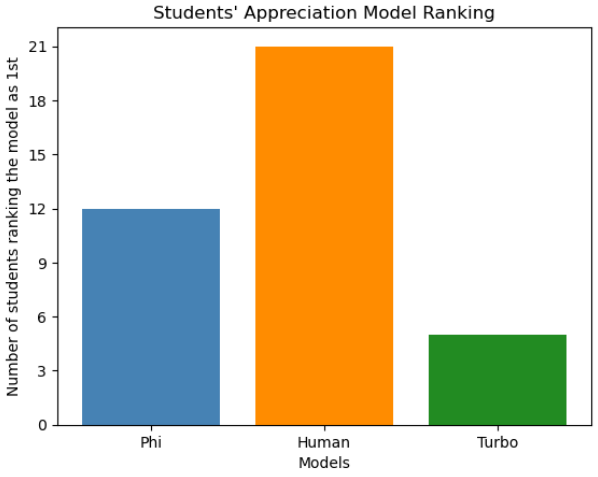


Figure 10: Students' Appreciation of the Models

5 Conclusion

The results of our experiments show that the stories and questions generated by our fine-tuned model, Phi, are comparable to those created by teachers in terms of clarity and adaptability to the students' comprehension level, both outperforming GPT.
In terms of enjoyment, there was no significant difference between the three models overall, although Phi and Human received slightly higher average daily scores than GPT. However, in the final feedback, the Human model was ranked highest by the majority, followed by Phi and GPT. This suggests room for improvement in making LLMs more engaging. One potential improvement is incorporating personalized information, such as students' names and school details, into the prompts to create more relatable and amusing stories, thereby increasing student enthusiasm and interest.
The perceived learning outcomes across the three models were not particularly high, with Phi performing slightly better than the others. This could be attributed to the brevity of the sessions, limited interaction between the students and the robot, and the decision to keep the stories short to maintain student focus. Besides, much of the story content focused on character development and setting descriptions, leaving less room for educational content. Future improvements could involve breaking the stories into smaller segments and increasing interaction between the robot and the students to enhance engagement and learning.

References

[1] Jeanne M. Hughes, Justina Oliveira, and Crystal Bickford. "The Power of Storytelling to Facilitate Human Connection and Learning". In: *Impact: The Journal of the Center for Interdisciplinary Teaching Learning* 11.2 (July 2022).

[2] Bingsheng Yao et al. *It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books*. 2022. arXiv: 2109.03423 [cs.CL]. URL: https://arxiv.org/abs/2109.03423.

[3] Zheng Zhang et al. "StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement". In: *CHI Conference on Human Factors in Computing Systems*. CHI '22. ACM, Apr. 2022, pp. 1–21. DOI: 10.1145/3491102.3517479. URL: http://dx.doi.org/10.1145/3491102.3517479.

[4] Nilüfer Atman Uslu, Gulay Öztüre Yavuz, and Yasemin Koçak Usluel. "A systematic review study on educational robotics and robots". en. In: *Interact. Learn. Environ.* 31.9 (Dec. 2023), pp. 5874–5898.

[5] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG]. URL: https://arxiv.org/abs/2305.14314.

[6] Enkelejda Kasneci et al. *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*. en. EdArXiv. Jan. 2023. DOI: 10.35542/osf.io/5er8f. URL: https://doi.org/10.35542%2Fosf.io%2F5er8f.

[7] Marah Abdin et al. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. 2024. arXiv: 2404.14219 [cs.CL]. URL: https://arxiv.org/abs/2404.14219.

[8] Loubna Ben Allal et al. *Cosmopedia*. 2024. URL: https://huggingface.co/datasets/HuggingFaceTB/cosmopedia.

[9] Maha Elgarf, Hanan Salam, and Christopher Peters. "Fostering children's creativity through LLM-driven storytelling with a social robot". en. In: *Front. Robot. AI* 11 (Dec. 2024), p. 1457429.

[10] Lyumanshan Ye et al. *Connection is All You Need: A Multimodal Human-AI Co-Creation Storytelling System to Support Children's Multi-Level Narrative Skills*. 2024. arXiv: 2405.06495 [cs.HC]. URL: https://arxiv.org/abs/2405.06495.

[11] Chao Zhang et al. "Mathemyths: Leveraging large language models to teach mathematical language through child-AI co-creative storytelling". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Vol. 33. Honolulu HI USA: ACM, May 2024, pp. 1–23.

A Appendix

A.1 Tasks for the Teachers

Create one story (5 paragraphs) for each one of the topics below, as well as 3 questions with the answers for each story:

1 - Understand that environmental changes can sometimes threaten living things. Learn to create and understand food chains, identifying roles like producers, predators, and prey. Develop skills in asking questions and using scientific methods to find answers. Children will recognize their current knowledge and identify what they need to learn about food chains, as well as generate questions and decide on ways to answer them.

2 - Recognize that environmental changes, such as fires and floods, can pose risks to living things. Children will understand how these changes impact life and consider how scientists use food chains to predict the effects of environmental changes.

3 - A story about how these changes in the environment can be compared to multi-cultural integration and diversity, that can prompt children to reflect on themselves and their relationships with the environment.

A.2 Pre-test Survey Questions

Q1: Have you ever interacted with a social robot?
1. Never
2. Once or twice
3. More than three times in my life
4. Every week

Q2: How much do you like social robots? (1-5 scale)

How much would you like to have a social robot to do the following activities: (1-5 scale)
Q3: Helping with your tasks/lessons in the classroom?
Q4: Helping with your lessons at home (homework)?
Q5: Playing with you during the play time?
Q6: Helping you with other tasks?

Q7: How much you think you and the robot can be friends?

Q8: Which one of the lessons you like the most?
Q9: Which one of these lessons you think the robot can help you learn the most?
Mathematics & Technologies — Languages & Literature — Social Sciences & Arts

Q10: How much you think the robot can help you be friends with your peers? (1-5 scale)

Q11: How much you think the robot can help you feel better in the classroom? (1-5 scale)

A.3 Daily Post-test Survey Questions

Q1: How much did you like this activity? (1-10 scale)
Q2: How much do you think you learned with this activity? (1-10 scale)
Q3: How much did you like the robot's questions and answers? (1-10 scale)
Q4: How much did you prefer this activity compared to the one done yesterday? (1-10 scale)

A.4 Final Feedback Survey Questions

Storytelling of food chains:
Rate on a 1-5 scale
What did you like about it?
What did you dislike about it?
Rank the model from 1 to 3

Storytelling of changes in the environment:
Rate on a 1-5 scale
What did you like about it?
What did you dislike about it?
Rank the model from 1 to 3

Storytelling of working together:
Rate on a 1-5 scale
What did you like about it?
What did you dislike about it?
Rank the model from 1 to 3

Additional Feedback
What would you tell the robot that you learned the most with it?
Why was the robot so important?