Math teaching through storytelling, LLMs or humans?

Ismail El Azhari
ismail.el.azhari@umontreal.ca
University of Montreal
Montreal, Quebec, Canada

Noé Jager
jager.noe@umontreal.ca
University of Montreal
Montreal, Quebec, Canada

Sparsha Mishra
sparsha.mishra@umontreal.ca
University of Montreal
Montreal, Quebec, Canada

ABSTRACT

Large Language Models (LLMs) are becoming increasingly integrated into educational settings across various age groups and affiliations. This raises an important question: How do their teaching capabilities compare to those of human teachers? This subject has been handled by previous research, by it seems like previous researchers mostly explored LLMs' effectiveness in teaching mathematics to children. Our experiment on the other hand extends the inquiry to university students examining their learning experiences with LLM-generated stories versus human-authored stories on more advanced mathematical concepts. We conducted an experiment with 16 participants (university students),each exposed to two stories explaining different mathematical concepts: (Gradient Bayesian Probability), with participants blinded to the authorship to eliminate bias. Preliminary findings indicate a noticeable improvement in understanding the gradient concept among participants who read the LLM-generated stories compared to the human-authored ones. These results suggest potential advantages of LLMs in teaching specific types of mathematical content, warranting further investigation into their role in higher education.

KEYWORDS

Large Language Models, Mathematics, Education, Storytelling, University Students, Human-Computer Interaction

ACM Reference Format:
Ismail El Azhari, Noé Jager, and Sparsha Mishra. 2024. Math teaching through storytelling, LLMs or humans?. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1. Problem space and why it matters (1 paragraph) 2. "Past work wasn't good enough / didn't address this exact problem" (1 paragraph) 3. What we did / our solution (1-2 paragraphs) 4. Tease major contributions/findings (could be a short bulleted list or single paragraph) 5. (Optional) A short paragraph overviewing the structure of the paper (esp. useful for longer papers or more complex structures)

The rapid advancement of large language models (LLMs) such as ChatGPT, Google Gemini, and Llama has brought significant changes to educational methodologies. These LLMs are increasingly recognized for their ability to present complex concepts in engaging and creative ways such as storytelling. However, the question remains: can LLM-generated stories effectively rival the insights, creativity, and pedagogical nuances of human teachers? Given the foundational role of mathematics in higher education and its importance in critical thinking, finding the best methods for conveying mathematical ideas to students is crucial.

While previous research, such as the "Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling" study or the "Using Children's Literature to Teach Mathematics: An Effective Vehicle in a STEM World" study by Joseph M. Furner from Florida Atlantic University, highlighted the potential of LLMs to teach math effectively through storytelling, these studies primarily focused on children and lacked a direct comparison to human educators in structured academic contexts. Moreover, it did not evaluate the nuanced aspects of learning, such as the appeal and understandability of the teaching medium for adult learners. This leaves a gap in understanding whether LLMs can replicate or even surpass human teachers' ability to convey abstract mathematical concepts like Bayesian reasoning or gradients to university students through the art of storytelling.

So in order to enlarge the scope of previous work in this field, our study compares the effectiveness of LLM-generated and human-written stories in teaching university-level math concepts. We focused on two topics, Bayesian reasoning and gradient, selected for their foundational importance in mathematics and their potential to challenge learners. Using a within-subject methodology, we recruited university students as participants. Each participant engaged with two stories, one generated by an LLM and another crafted by a math professor. To ensure rigor, we balanced the order of storyteller presentation among all our participants to control for potential biases caused by the story sequence. Data collection incorporated both qualitative methods; surveys and interviews, and quantitative measures; pre-test and post-test scores on exercises, to evaluate story appeal and understandability.

Our findings reveal intriguing insights into the comparative effectiveness of LLMs and human-authored stories in teaching mathematical concepts :

- **Baseline Equivalence:** Both groups of participants showed no prior differences in understanding the concepts before engaging with the stories, ensuring the validity of our comparisons.
- **Concept-Specific Outcomes:** While no significant differences were observed in the grades depending on the story author or concept, we did find a difference in improvement in understanding.
- **Interaction Trends:** We observed a trend suggesting that the relationship between the type of story and its effectiveness may depend on the mathematical concept.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **Implications for Educational AI:** These findings suggest that LLMs have a slight advantage in teaching certain mathematical concepts, pointing toward their potential as powerful educational tools when applied strategically.

These contributions pave the way for future research into how LLMs can be used more effectively to harness their full potential while integrating, rather than discarding, the unique strengths of human educators in teaching mathematics.

2 METHODOLOGY

Immediately below header: Overview of research hypotheses/questions, study design & justification

Subsection: Participants & Recruitment

- Describes who the participants were, demographic info, recruitment procedures & screening criteria

Subsection: Study Design, Tasks, and Procedure

- Our design and study protocol. Also mentions compensation, if any.

Subsection: Data collection, Setup, and Materials

- How and what data were collected, the specific software/hardware setup, the instrument(s) used (e.g., NASA TLX), if not already mentioned.

Subsection: Data Analysis (sometimes included)

Immediately below header: Overview of research hypotheses/questions, study design & justification

2.1 Participants & Recruitment

We recruited 16 participants who were university students, as they represented our target demographic. The participants ranged in age from 21 to 28. They were informed that the study aimed to evaluate the effectiveness of storytelling as a method for teaching mathematical concepts. Participation was voluntary, and all participants signed a consent form before taking part in the study. Instead of conducting a screening process, we opted to include all participants and retained both their pre-test and post-test scores to obtain a comprehensive view of their improvement based on the story author and subject.

2.2 Study Design, Tasks, and Procedure

Our study employed a within-subjects design to evaluate the effectiveness of storytelling by both human authors and large language models (LLMs) in teaching two mathematical concepts: Bayesian Probability and Gradient Descent. We collaborated with two math teachers and two LLMs (ChatGPT and LLaMA), each contributing one story for each mathematical concept. This setup created 16 unique orderings, ensuring that every participant experienced both concepts and story authors. Sixteen university students were recruited, with each participant assigned to one of the possible orderings.

Participants began the study by completing two exercises related to the mathematical concepts to establish a baseline of their prior knowledge. The likelihood of a participant answering all baseline questions correctly by chance was calculated as (1/4)*6=0.000244 or 0.024%, ensuring a robust measure of pre-existing knowledge.

Following the baseline assessment, participants read the first story and answered the same exercise questions to evaluate their understanding of the presented concept. They also completed a survey featuring Likert-scale questions designed to assess the story's appeal and understandability. The process was then repeated for the second story, with participants answering the exercise questions and completing a second survey. After completing both tasks, participants participated in a semi-structured interview aimed at capturing their overall impressions of storytelling as a method for teaching mathematics. The interview also explored whether participants suspected the true purpose of the study, thereby assessing potential biases. No monetary compensation was offered for participation. However, participants were thoroughly debriefed after the study, during which the true intent of the research was revealed. Even after the reveal, none of the participants indicated that they had suspected the true purpose of the study beforehand. This structured design ensured a comprehensive evaluation of how the story author and subject matter influenced learning outcomes while minimizing bias and maximizing methodological rigor.

2.3 Data collection, Setup, and Materials

To streamline the data collection process, we developed a dedicated website where participants completed all tasks, including answering the exercises, reading the stories, and filling out the surveys. This approach ensured consistency and accessibility across the study. The only exception was the final interview, where participant responses were manually transcribed by the researchers during live sessions.

The setup for the study was designed to minimize distractions and maintain a controlled environment. Some sessions were conducted in person, while others were facilitated online through Discord with screen sharing. In both cases, participants worked independently without access to any materials beyond the website. This controlled environment ensured that the results were not influenced by external factors and that participants fully focused on the tasks at hand.

Participant	Order	StoryScenario	StoryAuthor	PreResults	PostResults	Improvement	Median_Survey_Score
1	1	bayes	Human	2	4	2	5.0
1	2	gradient	LLM	4	5	1	6.0
2	1	bayes	Human	3	3	0	4.5
2	2	gradient	LLM	2	4	2	4.0
3	1	bayes	Human	2	3	1	4.0
3	2	gradient	LLM	0	4	4	5.5
4	1	bayes	Human	1	4	3	5.0
4	2	gradient	LLM	4	4	0	5.5
5	1	gradient	Human	5	2	-3	5.0
5	2	bayes	LLM	4	4	0	6.0

1–10 of 32 rows Previous 1 2 3 4 Next

Figure 1: Overview of the data structure after cleanup

2.4 Data analysis

Our data was categorized into two types: quantitative data and qualitative data. For the quantitative data, we used Google Sheets to initially record the scores of each participant. The data was then imported into R, where we transformed it into a long format. This step was crucial for managing within-subject data, as each participant interacted with two stories.

We began the analysis by visualizing the data. QQ plots and histograms were created to inspect the distribution of the scores. To formally test for normality, we applied the Shapiro-Wilk test. The

results indicated that the data did not follow a normal distribution, prompting us to test the residuals of our models for normality. For statistical analysis, we employed linear mixed-effects models using the (lmer from the lme4 package) to account for the repeated measures within participants. Homoscedasticity was assessed through spread plots, and Levene's test was used to confirm the findings. These steps ensured that the model assumptions were appropriately validated.

After validating the assumptions, we analyzed the effects of story author (LLMs vs. humans) and mathematical concepts (Bayesian vs. Gradient) on participant scores. Pairwise comparisons were conducted using the emmeans package to explore significant differences between conditions. This step helped us identify whether the interaction between the type of storyteller and the mathematical concept influenced learning outcomes.

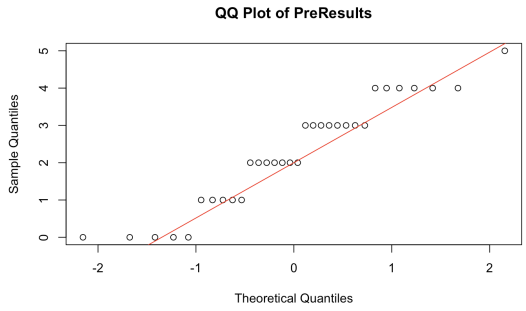

Figure 2: QQ plot for PreResults scores

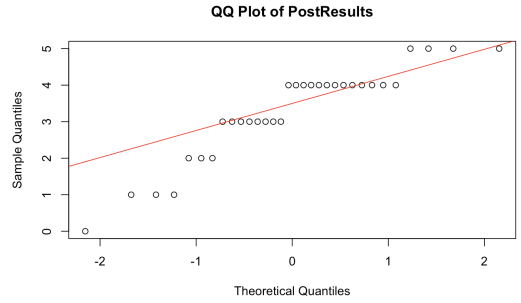

Figure 3: QQ plot for PostResults scores

3 FINDINGS:

3.1 Quantitative findings:

3.1.1 LLMs (Bayesian and Gradient).

The analysis revealed differences in the effectiveness of LLM-generated stories based on the mathematical concept being taught. Participants showed higher scores on average, for gradient stories compared to Bayesian stories. Post-hoc analysis using estimated marginal means showed that gradient stories generated by LLMs resulted in a higher average score (4.00) compared to Bayesian stories (3.38). However, while these differences are numerically notable, they did not reach statistical significance (p= 0.156 > 0.05). This

```
        Shapiro-Wilk normality test

data:  data_long$PreResults
W = 0.92563, p-value = 0.02964


        Shapiro-Wilk normality test

data:  data_long$PostResults
W = 0.88179, p-value = 0.002198
```

Figure 4: Shapiro-Wilk test results for both PreResult and PostResult scores

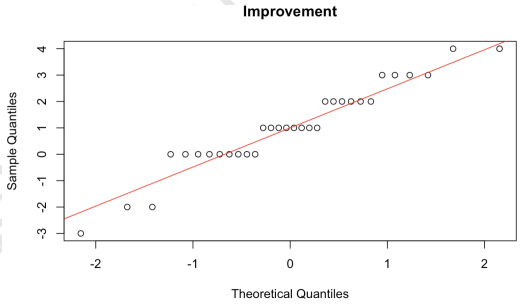

Figure 5: QQ plot for Improvement scores

```
        Shapiro-Wilk normality test

data:  data_long$Improvement
W = 0.94163, p-value = 0.08331
```

Figure 6: Shapiro-Wilk test results for Improvement scores

suggests that while LLMs may be particularly effective for teaching more visual or abstract mathematical concepts like gradients, the observed trend requires further investigation with a larger sample size to establish its robustness.

3.1.2 Humans (Bayesian and Gradient).

For human-authored stories, the results contrasted with those for LLMs. Bayesian stories outperformed gradient stories, with an average score of 3.50 compared to 2.25 for gradient stories. This aligns with the hypothesis that human-authored stories might be better suited for concepts requiring narrative structure or logical reasoning, such as Bayesian reasoning. Post-hoc tests confirmed that while Bayesian stories performed better than gradient stories in terms of participant scores, the differences were not statistically significant (p > 0.05). These findings suggest an apparent trend favoring human-authored stories for certain types of concepts. However, the variability in performance and the lack of statistical significance underscore the need for further research to confirm these patterns.

3.1.3 *Improvement (Bayesian and Gradient).*

Improvement scores were calculated as the difference between pre-test and post-test scores. For Bayesian stories, human-authored narratives demonstrated greater improvement (1.38) compared to LLM-authored ones (1.00). However, this difference was not statistically significant ($p = 0.6408$). In contrast, for gradient stories, LLM-authored stories demonstrated significantly greater improvement (1.75) compared to human-authored ones (0.00), with a p-value of 0.0365. These findings suggest that the effectiveness of the storyteller is influenced by the type of mathematical concept, with LLMs excelling at teaching visual or abstract concepts like gradients, while human-authored stories may be more effective for logical or narrative-driven concepts such as Bayesian reasoning.

3.2 Quanlitative findings:

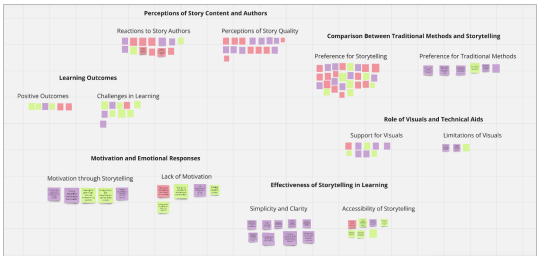

Figure 7: Affinity diagram using Miro

3.2.1 *Theme 1: Storytelling as a Tool for Simplification and Engagement:*

Participants widely acknowledged the effectiveness of storytelling in making mathematical concepts simpler, clearer, and more relatable. Storytelling was seen as a valuable introductory tool, particularly for those who struggle with traditional approaches. Several participants emphasized that stories connect abstract concepts to real-world applications, making them easier to understand and more engaging.

Examples:

- "Finding storytelling enhances understanding."
- "Connecting more with real-life situations and concepts."
- "Considering using stories to learn new math concepts again in the future because it helps him visualize the concepts in day-to-day life."

3.2.2 *Theme 2: Perceived Limitations of Storytelling for Advanced Learning:*

While storytelling was valued for its simplicity, participants frequently noted its limitations for deep or advanced learning. They believed storytelling lacked the depth needed to fully understand complex concepts or apply them in technical contexts. This was particularly evident when dealing with abstract or unrelated story backgrounds.

Examples:

- "Limited applicability of storytelling for advanced topics."
- "Thinking it is great to get a first idea, but not to learn the full concept."
- "Not understanding concepts because of unrelated backgrounds."

3.2.3 *Theme 3: The Role of Visual Aids in Enhancing Comprehension:*

Visuals emerged as a recurring suggestion to enhance storytelling's effectiveness. Participants highlighted that diagrams, graphs, or other visual representations could help clarify complex concepts, particularly in stories that involve mathematical details. However, some cautioned against overloading with visuals early on, emphasizing the need for a balanced approach.

Examples:

- "Thinking added visuals would be helpful."
- "Explaining visuals would have helped more."
- "Cautioning against overloading with visuals at the start."

3.2.4 *Theme 4: Mixed Reactions to Story Quality and Perceived Authors:*

Participants' reactions to the quality of the stories varied, with some finding certain stories easier and clearer than others. Interestingly, the perceived authorship of the stories (AI vs. human) influenced their evaluations. Participants were often surprised by the origin of the stories and cited issues like poor wording, lack of detail, or complexity as barriers to understanding.

Examples:

- "Concluding story 1 was written by a math teacher and story 2 by LLM."
- "Explaining poorly written stories hinder comprehension."
- "Feeling surprised the more confusing story was written by a teacher."

4 CONCLUSION:

4.0.1 *Summary of the significant findings of our results.*

Our study provided several important insights into the comparative effectiveness of the art of storytelling using human authored stories and large language model generated stories when teaching mathematical concepts to university students using. These are our some of our key findings:

LLMs Excelled at Visual/Abstract Concepts:

For the gradient concept, LLM-authored stories led to significantly greater improvement in participant understanding compared to human-authored ones ($p = 0.0365$), coincidentally, our participants in general were more inclined to compliment the gradient story if it was written by an LLM. This suggests that LLMs may be particularly effective for teaching abstract concepts in mathematics.

Humans Outperformed in Narrative-Driven Concepts:

It might not be time just yet to retire our human teachers, as we did find that humans outperformed LLMs in the Bayesien probability stories, demonstrating greater improvement scores, albeit not reaching statistical significance ($p = 0.6408$). This is perhaps something that will need further research using different narrative driven concepts and with larger pools of participants.

Trust in human written stories over AI written ones:

This is perhaps the finding that made us the most glad that we conducted a deception study. During our interviews, when we asked our participants to guess which stories was written by a human teacher, their guesses did not yield any significant findings. However, the reasons they gave to justify their guesses were very revealing, stating that they guessed the stories they found to be

well written and easier to understand to be the ones written by teachers, even though for the gradient decent stories that was not the case at all. This shows that we inherently have this bias against AIs believing that they will weave complicated tales due to the complexity due to complexity in understanding how they work, we believe that a human teacher will relate to us better and therefore make the story easier to understand.

4.0.2 *Validity threats:*

1st Obstacle: Study Environment

Conducting the study in varied settings (classrooms, library rooms, or online) introduces environmental variability that might affect participants' focus, comprehension, or engagement. Online participants may face distractions or technical issues not encountered in controlled physical environments. These inconsistencies could lead to differences in how participants experience the storytelling task, potentially influencing their responses and compromising internal validity.

2nd Obstacle: Language Barriers

Limited English proficiency may affect participants' comprehension of the stories, influencing their understanding and engagement. This introduces a construct validity threat since the measurements (e.g., understanding, appeal) may not accurately reflect their ability to grasp the concepts. It also limits external validity since findings might not generalize well to populations with varying language proficiency levels.

3rd Obstacle: Technical difficulties

Technical issues (e.g., server errors, and inaccessible materials) could degrade the participant experience, disrupt data collection, or cause participants to disengage. This introduces an internal validity threat, as it could skew the results or lead to participant dropout. Technical issues also compromise reproducibility, as they create challenges for replicating the study in similar conditions.

4.0.3 *Putting the study in the body of work:*

The question of whether a machine can ever exceed human capabilities in fields that require adaptability and an understanding of human emotions is a complex one. While the answer has eluded us for decades, recent advances in large language models (LLMs) bring us closer than ever to finding it.

The father of this field of research, Mr. Alan Turing, once said: "A computer would deserve to be called intelligent if it could deceive a human into believing that it was human." Today, the question is: Are our computers already worthy of being called intelligent?

In our case, we focused on investigating a very specific aspect of this perceived intelligence: the art of storytelling while teaching mathematical concepts. This inspiration stemmed from two papers that explore teaching children mathematics through stories written by LLMs:

"Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child AI Co-Creative Storytelling" "Using Children's Literature to Teach Mathematics: An Effective Vehicle in a STEM World." These papers laid the foundation for our hypothesis. Initially, our inquisitive yet skeptical minds believed that such findings could only apply to children's stories. We assumed that the novelty of an AI-written story might be so exciting for children that it could capture their attention and lead to better retention of information. However, our study proved us wrong—and perhaps, the dawn of AI teachers is indeed upon us.

That said, we recognize the limitations of our study, both in its scope and its reach. We hope that further research in this field will build upon our findings, expanding our understanding of AI's teaching capabilities and its potential applications.

REFERENCES

A RESEARCH METHODS

We conducted a within-subject usability study to compare participants' learning outcomes when reading AI-generated and human-written stories.

This is a mixed method study relying on both qualitative and quantitative data.

Independent Variable (IV): StoryAuthor (Human vs LLM), StoryScenario (Bayes probability vs Gradient descent), Order (1st vs 2nd)

Dependent Variable (DV): PreResults (Exercice score before engaging with the story), PostResults (Exercice score after engaging with the story), Median_Survey_Score (Median score of 7 survey questions after reading the story).

B ONLINE RESOURCES

Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling

Using Children's Literature to Teach Mathematics: An Effective Vehicle in a STEM World