# Math teaching through storytelling, LLM, or Humans?

IFT 6075 - Empirical Methods in HCI
Initial Project Proposal

From the worldly renowned Triad Matrix, also known as

Ismail EL AZHARI
Sparsha MISHRA
Noé JAGER

## Project Abstract

With the rapid development of LLMS like Chatgpt, their role in education is gaining significant attention. As we see these models getting increasingly advanced these days, debates have emerged about whether LLMs can rival or even replace traditional teachers in the classroom.

Inspired by the article Mathemyths, which argued that LLMs could teach math to children more effectively than teachers, our research hypothesis challenges this view. We believe that while LLMs may offer a novel and engaging way of teaching, their capacity to truly enhance learning outcomes, particularly in math, remains questionable when they are compared to insights, intuition, and teaching methods that traditional teachers bring to the table.

To get to the bottom of our hypothesis, we have chosen to conduct a study using a population of university students who will be presented with two stories about two different mathematical concepts. Our participants will be unaware of the nature of the writers and will only discover it after they are done with the process. We are using a mixed method by splitting our participants into two groups and giving them a different experience when it comes to the ordering of the story maker. At the same time, each participant will have two stories and will experience both being taught by an LLM and by a human teacher.

## Project Goals and Research Hypotheses

We want to see if **LLM applications** are **better** than **math teachers** at **teaching math concepts**. To follow the path of the MatheMyths article, we add another constraint that is they **must use storytelling** to teach the participants. Hence, comes our research hypothesis:

> "When it comes to teaching mathematical concepts through the art of storytelling, using a text LLM or a professional teacher to write the story, has no impact on the level of appeal and understanding of the concept by an adult university student who is not familiar with the concept."

## Study Design

### • What methods will you use and why? (E.g., between-subjects, within-subjects, or a more complicated design?)

We will be using a **mixed method incorporating both between-subjects and within-subjects** methods. The reason for that is that we need to test all the participants on both storytellers to have **more control over the outliers** among our participants, making it a within-subjects study and making us test on two math concepts.

However, doing so forces us to deal in some way with the **ordering**, that maybe concept 1 would be done by AI and that it would be so much harder than the other one done by the other, influencing the results. Hence, we do so by **separating the participants into two groups** and teaching them concepts one and two with both LLM or Maths teachers and inverting who teaches what.

Due to the risk that participants might already know the concepts (or one of them), we will conduct a **pre-test screening** to eliminate those who are familiar with the concepts.

Additionally,

- To keep a good **construct validity**, we will **generate stories from three different math professors** — knowing their domain, and used to teach those concepts — hence removing the doubt on the quality of the human stories.
- To help with the **reproducibility of our results**, we will also **generate stories from 3 distinct LLMs** (ChatGPT, Google Gemini, and Microsoft Copilot). A full reproducibility isn't possible since every story generated by one of those applications is partially randomized.

(Both randomly and equally distributed, their stories among the participants.)

Here is a simplified version of the two groups of participants to help you understand better:

- Pre-test screening
- Group 1:
    1. Users learn **Concept 1** with one of the **LLM** story + Exercise
    2. Users learn **Concept 2** with one of the **Math professor**'s story + Exercise
- Group 2:
    1. Users learn **Concept 1** with one of the **math professor**'s story + Exercise
    2. Users learn **Concept 2** with one of the **LLM** story + Exercise

## • Who will participate and why? How will they be recruited?

This study requires us to have **math professors** as well as **students as participants**. We plan on recruiting them via our relations (friends, family members, coworkers) and under the pretext of participating in a research experiment from the prestigious University of Montreal.

### • How many participants will there be and why?

We aim at gathering a population of **30 students** which we will **equally divide among the two groups.** We estimate — for the experience we want to conduct, for our capacities, and the fact that this is a university class project — that this is a fair and sufficient objective.

The **constraints of time and money** do not allow for a very large number of participants, however, we want to make sure that we have enough participants for our study to be representative of the adult population we chose as a subject.

### • How long will the study take for one participant?

- Each participant will have an introduction to tell them about what they will be doing today so they are ready. No information about who created the stories or the existence of other groups will be shared. ~ 5min
- A screening test, to test their knowledge of the two concepts that they will see. ~ 8 min
- Stories. ~ 16min (~ 8min *2)
- Exercises with Think Aloud. ~ 20 min (~ 10min *2)
- We finish with an interview and a survey. ~ 10min
- **Total: ~ 1 hour (~=59min)**

- **Estimated cost of study, in terms of time and money?**

The study is conducted on a **voluntary basis**; we have not considered offering any monetary compensation to the participants.

## • What are the independent and dependent variables?

IV - Storytellers (LLMs / Maths Teachers)

DV - Story Appeal, Story Understandability

## • What are the measures (how will dependent variables be measured)

1. Appeal (Engagement, Creativity) - ("How creative did you find the story?")
2. Understandability - Likert Scale from the survey (how clearly did the story explain the math concept? 1 = not clear, 5= very clear), Think aloud and results of the exercises

## • How will the data be collected?

**Qualitative data** will be gathered through the think-aloud method, where participants articulate their thoughts while solving exercises. Fieldnotes will document key moments, and interviews will provide additional insights into their experiences.

**Quantitative data** will be collected through post-task surveys and participants' performance on exercises.

- **Qualitative data:** Interviews, Think Aloud (+ Fieldnotes)
- **Quantitative data:** Surveys, Results in the exercises

## • Threats to validity:

- **External:**

Are the stories produced by Storytellers really representative of LLMs and humans?
Selecting three teachers at random leaves us open to the possibility that we have not selected the best teachers capable of combining the art of storytelling with their math teaching skills. The same thing can be said of the LLMs, especially as more and more of them are being developed in recent years.

    - **Reproducibility:**

      Full reproducibility isn't possible due to a few factors.

      1. Since every story generated by a Large Language Model application is partially randomized, it is not possible to obtain the same story even if the input is similar.
      2. The same applies to math teachers since their stories are dependent on their creativity, writing skills, and mindset at the moment of writing the story.
      3. The qualitative data collected from the Think Alouds and the interviews with our participants might vary among other populations.

         Hence, there is a chance our results might vary even using the same LLMs, teachers, and topics and/or participants are used.

- **Ecological validity:**

People are accustomed to certain learning methods that might impact our study.
Cultural differences can also make certain stories more relatable or easier to understand than others, so a cultural factor may play a slight role in relating to the stories.

- **Fatigue:**

After doing the pre-selection test and doing the first exercise, it is possible that the participant might get tired and be less focused when reading the second story, or perhaps take longer than allowed to understand it. Therefore, it is possible that fatigue might skew the results of whichever story comes second.