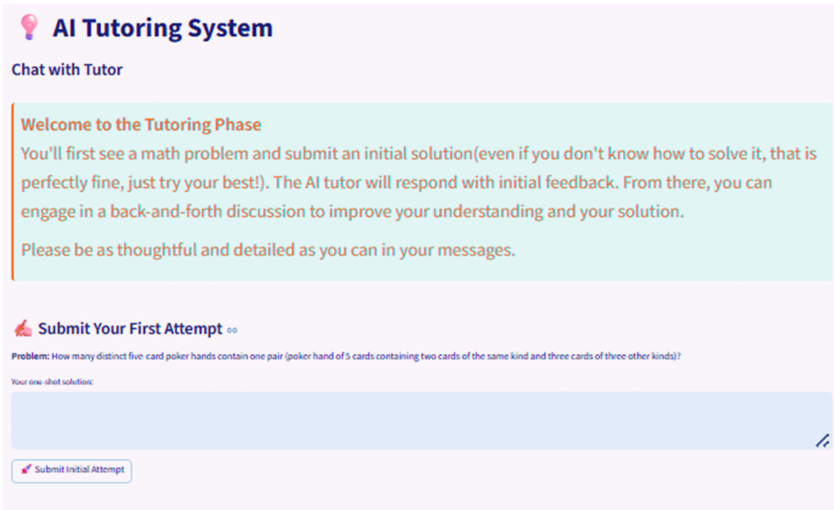
Pedagogical ChatBot: Designing Interactive FeedBack using Large Language Model(LLM)

Bianca Pitu - Computer Science, ÉPFL
Supervisors: Fares Fawzi, Tatjana Nazaretsky, Tanja Käser

Learning discrete mathematics can feel like climbing a mountain without a guide. What if students had an intelligent companion that offered timely, constructive feedback at every step?



1.Data Preparation

From 2,534 EdForum AICC1 posts(2022 - 2024), we used a hybrid filtering system to detect genuine student trial solutions.

- ***Filtering*** - Posts were scored with rules (keywords like "I tried to solve", "my proof", "j'ai essayé") and semantic similarity to example phrases, points added for reasoning markers ("therefore", "let … =") and subtracted for administrative content.
- ***Augmentation*** - High-scoring posts augmented with their original homework statements and official solutions

2534 → (Filtering) 138 → (Augmentation) 41

2. Simulating interactions between Student and Tutor Agents

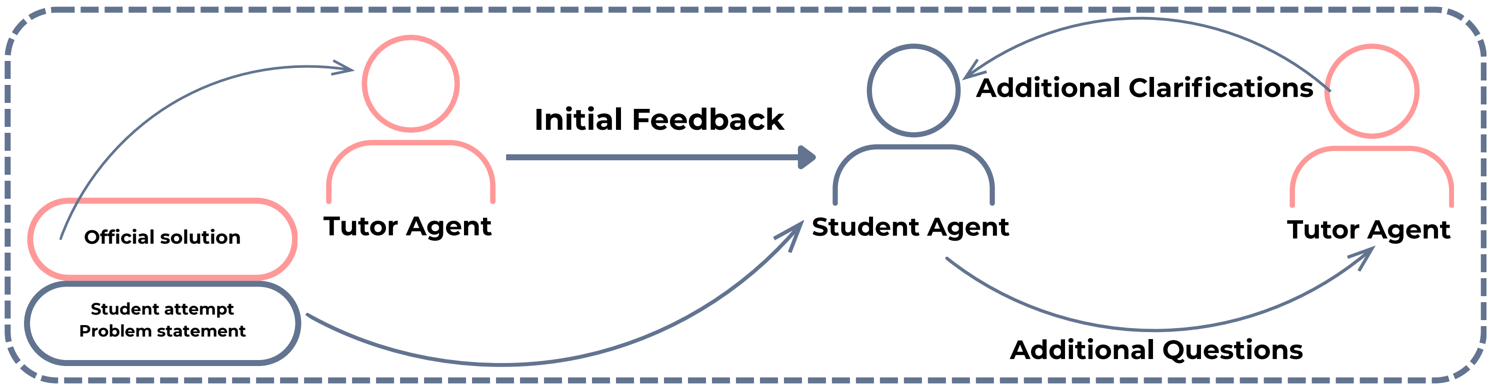
We simulated ***student–tutor interactions*** using GPT-4o agents to test different feedback strategies, **strenghts** and **weaknesses** before running a user study.

Tutor Agent: "Start by assuming n^2 is odd... As n^2 is a square of a number, it can also be written as n×n . So, n×n=2k+1. However, if n×n were even …"

Verbosity without clarity adds little new information

Scaffolding guides learner with focused reflective question

Tutor Agent: "Ask yourself: what do you assume in contraposition compared to contradiction?"



Strenghts	Weaknesses
Adaptive Explanations shifted from abstract definitions to concrete examples when needed.	**Verbose but Shallow** feedback was often long without adding meaningful insight.
Positive Reinforcement highlighted what was correct before addressing mistakes.	**Missed Misconceptions** sometimes failed to address the learner's actual difficulty.
Scaffolding asked guiding questions and broke reasoning into smaller steps.	**Uneven Coverage** explained some errors thoroughly while glossing over others.

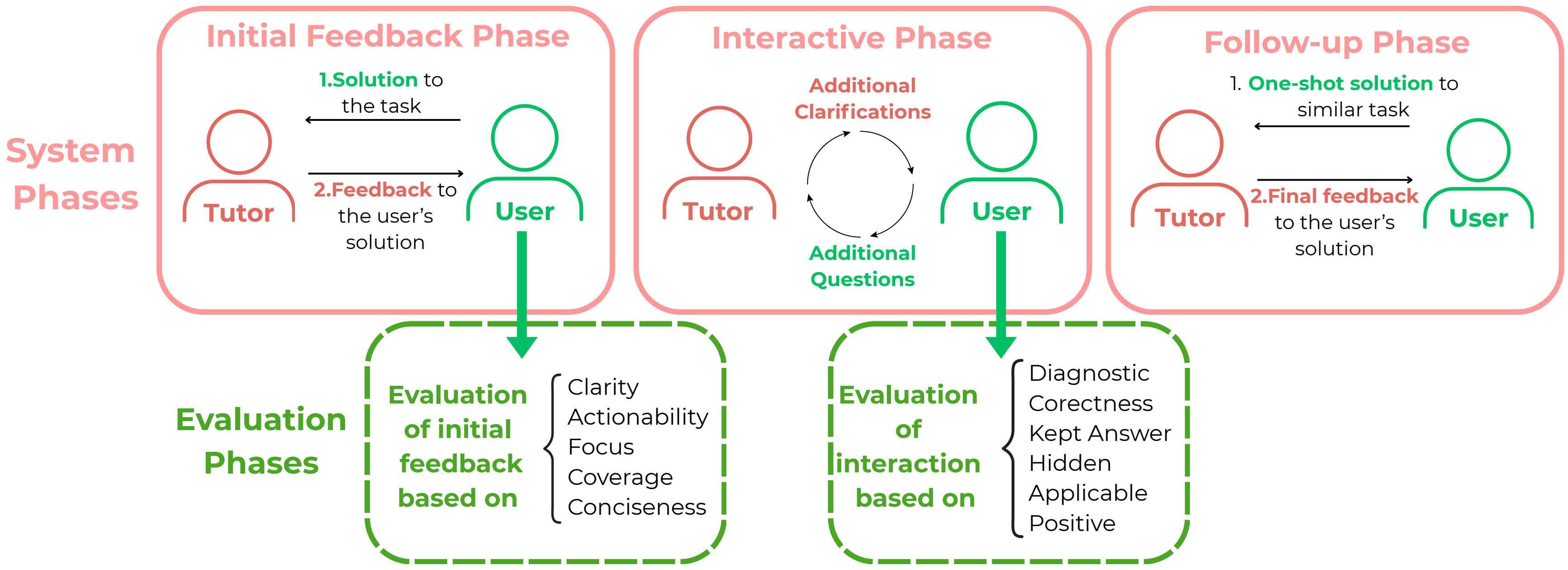
3. User Study Design

We built a **user-friendly web interface** powered by a **pedagogically prompted GPT-4o model** that resembled a tutoring app

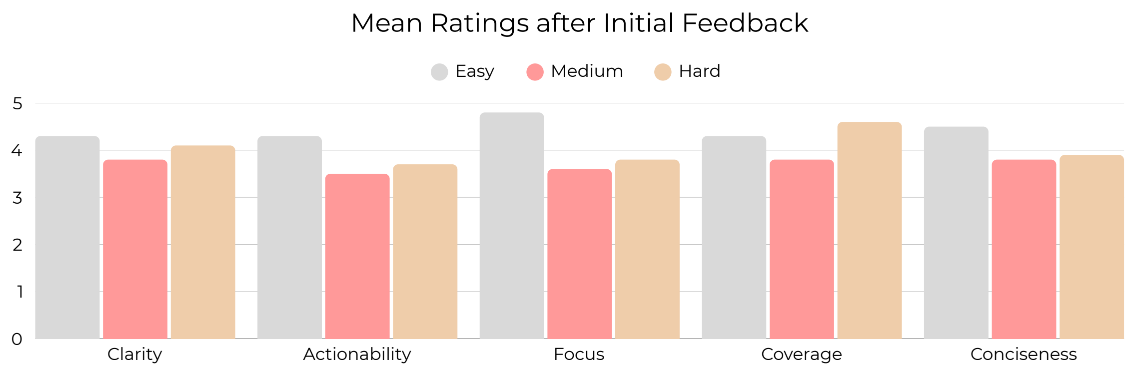
- three discrete mathematics tasks(Easy, Medium, Hard)
- 25 participants

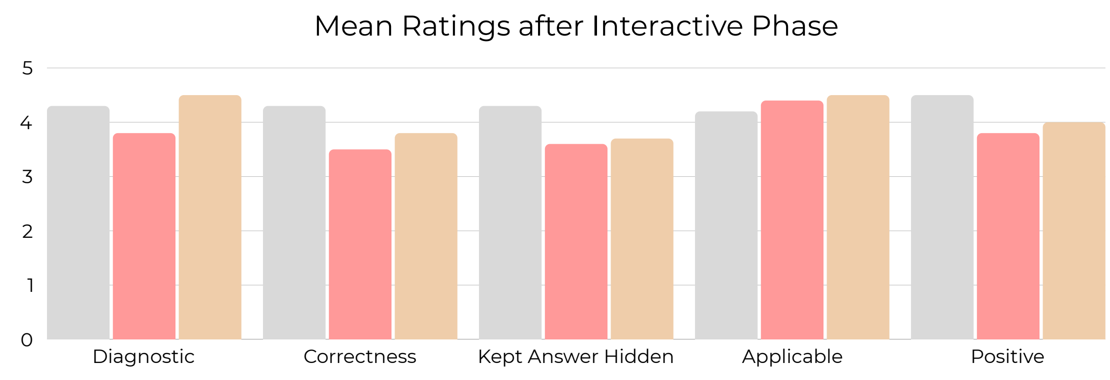
To address weaknesses identified in the simulations, we refined the tutor prompts:

- ***Conciseness enforced*** – initial feedback capped at ~120 words.
- ***No spoilers*** – hints and analogies only, never direct formulas.
- ***Progressive scaffolding*** – solution revealed gradually



4. User Study Rubrics Ratings





5. Future Work

- **Reduce Verbosity, Increase Precision**
  Explore techniques (e.g., reinforcement learning with human feedback) to make feedback more concise, preserving clarity
- **Improve Misconception Diagnosis**
  Train or fine-tune models on annotated student data to better identify where reasoning goes wrong
- **Tone Calibration**
  Adjust prompts or fine-tune to prevent feedback from sounding harsh, especially on follow-up attempts.

6. Acknowledgements

This study was supported by the **Laidlaw Scholars Leadership and Research Program** internship and conducted at the **Machine Learning for Education** (ML4ED) Laboratory, EPFL. I gratefully acknowledge their support.