“Walk It Out”: An Embodied and Mobile AI Tutor for STEM Education

Johnson-Glenberg, M. C., Chaudhari, F. V., Patel, S. J., Baia, S., Vieyra, R., O’Brien, D., & Megowan-Romanowicz, C.

Embodied Learning
Embodied STEM
Instructional AI
LLM AI Tutors
Mobile Learning
Socratic Tutors

Interpreting motion graphs can be challenging for students. Classroom teachers lack the capacity to provide rapid, personalized feedback for each student during mobile learning activities. We created an embodied smartphone app using LiDAR technology that has been shown to significantly improve kinematics understanding. However, the app's current text-based feedback can be enhanced with an AI tutor that gives more adaptive and "Socratic" (querying and non-directive) feedback. This article describes the design process for creating a Socratic-style mobile AI tutor and our Retrieval Augmented Generation (RAG) LLM architecture. The RAG pulls from the authors' published base of physics education articles to ground the model. Human experts determined that Claude Sonnet 4.5 provided significantly more precise, reliable, and Socratic-style feedback compared to other LLMs. This team of learning engineers and physics teachers is now creating a semantic benchmark test to assess the quality of the LLM's feedback.

Many students struggle with making and interpreting motion graphs (Börner et al., 2015; 2019); Börner et al. describe the broader challenges in data-visualization literacy across many age groups. Our research group has created a mobile app that helps students interpret graphs in an embodied manner by "walking the graph" (Megowan-Romanowicz et al., 2022). Mobile learning (Sharples & Pea, 2014) has grown rapidly with the widespread classroom use of smartphones and tablets. Although sonic motion sensors have been used by physics teachers since the late 1980s (Brasell, 1987), the corrective feedback is usually not adaptive nor personalized. The lessons are often not embodied and the feedback is usually restricted to students self-interpreting how well their motion visually matched a pre-determined motion graph.

Newer phone and tablet devices enable learning to occur "anywhere, anytime", and use built-in sensors to capture real-world physical actions, such as motion or precise location (Huang & Chiu, 2015; O'Brien, 2021). Emerging AI-driven educational apps aim to address the issue of giving more personalized feedback, but their performance is inconsistent, particularly in real time (Guo et al., 2025; Kestin, 2025). There remains a clear need for Intelligent Tutoring Systems (ITS) that combine mobile sensor data with adaptive, embodied, and pedagogically sound AI feedback to enhance students' ability to make sense of graphical data representations. To our knowledge, no prior research has combined mobile LiDAR-based motion tracking with real-time AI feedback generated by large language models (LLMs) on graphing tasks. The current app is called Motion Visualizer (MV); it uses students' real-time locomotion (walking) to create immediate position-time graphs on a LiDAR-enabled smartphone or tablet. Our research questions center around creating an AI tutor and assessing the efficacy of the tutor designed to give Socratic-style, non-directive feedback. The work adheres to best practices in learning engineering (Goodell & Kolodner, 2023). That is, we use team-based designs, iterative processes, and team-based refinement consensus to ensure the AI feedback is concise, accurate, and helpful for students.

Research Questions:

RQ1) Using numerical inputs (embodied position and time locomotion data) and students' think-aloud language, is it possible for a symbolic language-based LLM to generate feedback responses that are accurate and helpful?

RQ2) How can the LLM prompts be optimized to support some of the best practices of pedagogy, i.e., Socratic questioning and adaptivity to students' ongoing errors?

RQ3) Is the field in need of new semantic benchmarks for evaluating AI tutor's graphing feedback?

Methods

The App

The Motion Visualizer (MV) app is one within a suite of apps in the Physics Toolbox Sensor Suite mobile application (Vieyra, n.d.). The MV tool was created with the support of a National Science Foundation grant and was launched to the public in 2023. The app was designed to integrate principles of embodiment and locomotion into STEM education. It uses Light Detection and Ranging (LiDAR) to calculate the distance from a stationary surface, such as a wall, to the phone. On the device screen, the student's position is plotted every 100 milliseconds, i.e., how far the learner is from the wall or stationary surface. As learners move toward and away from the wall, they can see their motion plotted on a graph almost instantaneously. See Figure 1 for a screenshot of a successful attempt at matching the purple target curve. When learners match their motion plot to the purple curve with an MSE of less than .10, they are rewarded with digital confetti.

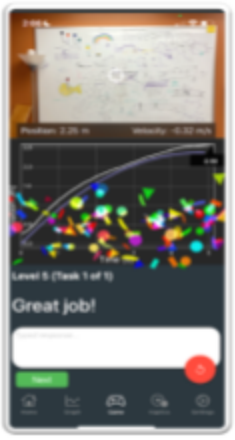

Figure 1. Screenshot.

Figure 2 shows an example of a failed attempt to match the purple target curve. The white dotted curve represents the student's locomotion. The non-AI version of MV gives only two types of text-based feedback: "Great Job" or "Try again". However, in our observations of teachers and peers supporting each other as learners, we noted the kind of feedback the students most benefited from was more nuanced. Effective feedback included suggestions about focusing attention on the starting position, considering how the target graph curves, and how the target graph shape might suggest how the learners' motions must change over time. Ultimately, the best kind of support helped students to not only build intuition about the meaning of position-time graphs, but to explicitly reflect on what the graph was telling them about their motion.

Data to test the AI models was gathered during a recent study

(Johnson-Glenberg et al., in preparation); the test set includes 90 transcripts of individual students' think-alouds and logfiles of the devices' "x, y" position-time coordinates across 14 graphing tasks. The tasks increased in difficulty over time; the final task was to recreate and match a vertical up parabola over a duration of 10 seconds.

What does it mean to be Socratic?

A difficulty in creating a pedagogically sound AI tutor is that LLMs have been pre-trained, hence predisposed, to give students the correct answer immediately. One of our goals was to create a feedback-giving system that would prompt the human user with a series of thought-provoking questions to encourage deep learning. The Socratic method of instruction (Paul & Elder,

2007) focuses on guiding learners to discover answers themselves via deduction and through the process of completing 14 graph-matching tasks. Directive feedback was given when necessary.

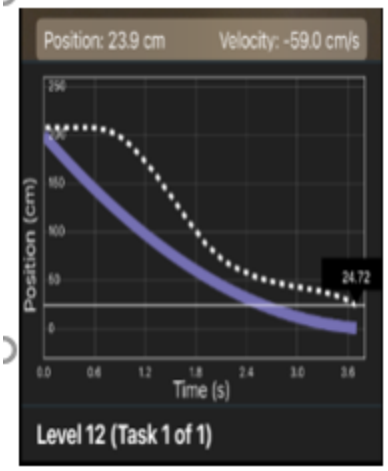

Figure 2. *Level 12, the purple line is the line the learners need to walk and match.*

Comparative LLMs

We probed three popular LLMs: Claude Sonnet 3.7, Claude Sonnet 4.5, and Meta Llama 3.1 70B, using the same prompts and RAG. Three human experts analyzed the results and came to a consensus that the most precise, consistent, and concise results were achieved by Claude Sonnet Model 4.5. Thus, Sonnet 4.5 was chosen as the primary model. This exercise alerted us that the field needs a new type of automated benchmark, one that is sensitive to physics feedback that will take into account semantics and the embodiment that goes into walking and creating digitized graphs (an admittedly new field of study). Updates on the new benchmark creation using an SLM and vectorized idea units are forthcoming.

Our AI Model

The goal now was to optimize the Sonnet Model 4.5 into an embodied Socratic Feedback Model. There are two levels of prompts. The first is called the global prompt, which sets the overarching principles that the AI system should adhere to for feedback across all responses. Our global prompt clearly identifies the role for the LLM: "...act as a professor of physics who adopts the Socratic method of instruction by using guiding questions that do not provide direct results of the solution to the problem faced by the learner during the first seven attempts". The consensus among two of the physics teachers in our research team was that they, as human tutors, would only give a student more directive feedback after the student had failed seven times. Based on their experience observing students, they would give explicit feedback only after failure seven times in a row, as students often began to show signs of frustration after that. The global prompt also defines the various forms of common mistakes that learners might make during the tasks. These common errors included having the wrong starting point of the graph and producing graphs whose slopes were too shallow or too steep (greater than .10 MSE from target line). After analyzing too many chatty and wordy LLM feedback responses (e.g., the early Claude responses always started with, "I noticed that…"), the training team decided to constrain the AI system to not give feedback that would surpass a 35-word limit. This also takes into account the UI space constraints on a smartphone.

In a parallel fashion, the system incorporated level-specific prompts that further refined the global prompt. "Level-specific" means the feedback was yoked to each unique instance of the 14 graphs. These level prompts enabled the LLM to make decisions on whether to 1) prioritize start position state, 2) the dynamics of the slope, 3) changes in direction, or 4) learners' acceleration patterns. The more advanced the task levels, the more complex the model's reasoning became, given that it must simultaneously incorporate Socratic questioning. Figure 3 shows the architecture of the AI model. The LLM tracks the number of errors, and, after the 7th attempt, the feedback becomes more directive. This was achieved by including the following language in the global prompt: (Note: Capitalization increases the likelihood that a Claude model will give special emphasis to the instructions.)

> "You guide students in a SOCRATIC manner. This means you ask questions until they have made 7 attempts. On the 8th attempt, you can start to give the students explicit answers that do not have a QUESTION MARK, but tell them exactly what they are doing wrong. The students' goal is to match the TARGET LINE."

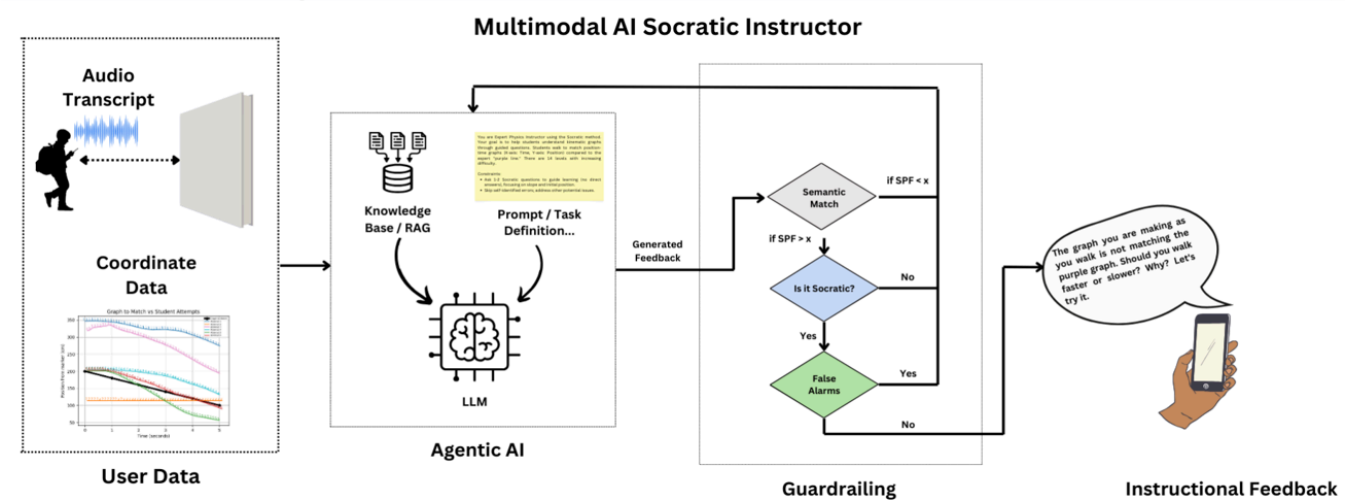

Figure. 3. Architecture of our Socratic AI Model.

Retrieval-Augmented Generation

To ensure that feedback is informed by studies on physics education, we created a lightweight retrieval augmented generation (RAG) module comprised of three papers on motion graph learning, education using smartphone sensor data, and previous studies from the design team (O'Brien et al., 2023; Vieyra et al., 2023; Vieyra et al., 2024). The RAG broke the papers into sections, and created a small vector index into which the sections, or chunks, were placed. For every attempt by the students, the RAG identifies types of errors and searches for extracted passages in the articles to inform its output. The passages are not directly exposed to or visible to the students, but serve as context for us, the testers, to ensure outputs are informed by solid kinematics research.

Results

To analyze the quality of feedback produced by the AI system, we assessed a corpus of 660 student attempts to match the graph across all 14 of the available tasks in MV. For each attempt, we determined if the output of the AI system satisfied the following criteria, including if it: 1) pointed to the correct errors, 2) adhered to the constraints of the Socratic querying syntax, and 3) suggested fixes that corresponded to the student-initiated locomotion. Below are examples of correct and incorrect feedback:

Correct Feedback Example: Figure 4 illustrates the student's attempt at Level 9. The initial discrepancy in this attempt is with the learner's start position.

Model Feedback Attempt #1: "Where should you be at the start so your line begins near the target line?

After multiple attempts, the student is now on their fourth attempt (Figure 4), and the model provides the following feedback.

Model Feedback #4: "Note your starting position. Did you start too far or close? Should you speed up in the beginning or slow down to better match the purple target line?"

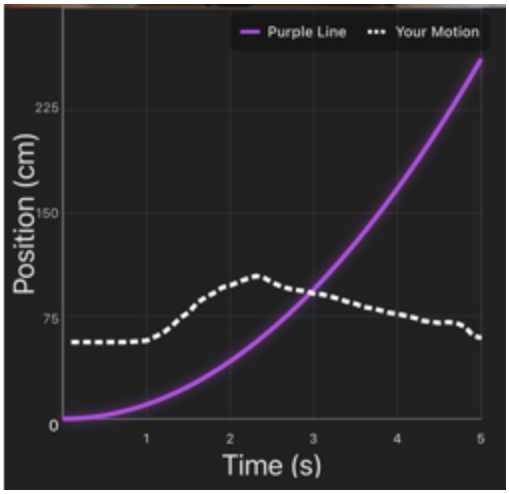

Figure 4. Attempt on level 9.

This feedback was considered effective as it maintains the required order of error-priority within the system, first nudging the learner to think about their mistake in the initial position, then moving towards analyzing higher-order thinking concerning adjusting the slope.

Incorrect Feedback Example: In an earlier training epoch, the AI feedback was:

Model Feedback Attempt #2.“Your line stays above the target line for most of the walk. What if you started moving faster earlier? Would that help your line drop to match the target's steeper curve?”

This feedback is incorrect as it indicates that the student's line is consistently above the target curve, and moving faster would not decrease the slope or make it “drop”.

F1 Score Evaluation

To provide an initial quantitative check on model performance, an F1 score was calculated using 50 randomly selected feedback samples. Using our classifications of each response as appropriate or not, the model achieved an F1 score of 0.81, suggesting strong alignment with the intended error-detection and feedback goals. See a confusion matrix in Figure 5. We are currently analyzing the responses within the dark blue cell (correct rejections) with a sensitivity analysis on the type of error and appropriateness of response with a new pilot semantic benchmark.

Discussion

We were consistently surprised by the time it took to train the LLM models to be precise, non-directive, and concise. We found that transcribing students’ vocal performance of their think-aloud protocols and then adding those transcripts as inputs to the LLM model was not helpful. This is because students say phrases like, “I need to go up”. The non-embodied LLM does not know the student is thinking of himself/herself as the line itself. It does not understand transmogrification or bodily metaphors.

Even after withholding the verbal inputs and feeding it only the position-time data, the models still ‘hallucinated’ on occasion and made incorrect analogies e.g., “imagine you are a ball rolling down a hill”. This type of AI-tutor feedback is especially egregious because one persistent error in understanding graphs is that naive human learners treat graphs like pictures (Beichner, 1994). After 12 months, and numerous LLM upgrades, the acceptability proportion of the output is now close to .81. The end goal is to encourage back-and-forth, multi-turn discourse between the tutor and the student. We are closer to that now that the model has been trained to remember and integrate the student’s previous trial performance (n = 3). We are examining recursion for memory using a model based on hidden units and working memory (Johnson-Glenberg, 2000). Our new system creates a lightweight memory for each student by storing key features of every attempt, including specific errors detected, slope issues, and starting-point mismatches. When a student submits a new trial, the model retrieves the last three attempts, summarizes common patterns of mistakes, and places this summary at the top of the level prompt used to generate the new feedback.

Future and Conclusions. We have created a Socratic AI tutor for a highly embodied graphing task with a LiDAR-enhanced mobile device. The model is adequate in theory, based on our evaluation of the model’s output, but still needs refinement and optimization. We will start to test its efficacy in the next semester with college students. Some researchers argue that because the current LLMs are trained solely on symbolic text that is inherently disembodied (i.e., culled from webpages and not experiential) and the models do not include perception, that the models will continue to make incorrect assumptions and never be able to ever fully “reason”. Rodriquez (2025) argues, “…that the path to genuine AGI necessitates a foundational shift towards embodied experience, proposing that intelligence is inextricably linked to the perceptual and action-oriented constraints of a physical body interacting with a persistent world.” We concur. The plan is to push on the boundary conditions

using symbolic text to help understand how to give relevant feedback for our highly embodied locomotive task. The next iteration may include perceptual inputs, something others have argued for, although we remind readers that using the think-alouds led the models to sometimes go astray. Thus, being more multi-modal did not help in this constrained physics task. The Socratic AI Tutor is currently running with AWS. A newly proposed upcoming research study will ascertain learning gains between the student group that uses the non-AI format of feedback in place now compared to the group that uses the Socratic AI tutor.

Acknowledgement

This material is based upon work supported by the National Science Foundation grants 2114586 and NSF NAIRR Supplement 2439960.

References

Beichner, R.J. (1994). Testing Student Interpretation of graphs. Am.. J. Phys. 62, 750–762 https://doi.org/10.1119/1.17449

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. Proceedings of the National Academy of Sciences, 116(6), 1857–1864. https://doi.org/10.1073/pnas.1807180116

Brasell, H. 1987. "The Effect of Real-Time Laboratory Graphing on Learning Graphic Representations of Distance and Velocity." Journal of Research in Science Teaching, 24 (4): 385–395. https://doi.org/10.1002/tea.3660240409.

Goodell, J. & Kolodner, J. (2023). Learning engineering toolkit: Evidence-based practices from the learning sciences, instructional design, and beyond. Routledge. https://library.oapen.org/handle/20.500.12657/86250

Guo, S., Halim, H. B. A., & Saad, M. R. B. M. (2025). Leveraging AI-enabled mobile learning platforms to enhance the effectiveness of English teaching in universities. Scientific Reports, 15(1), 15873. https://doi.org/10.1038/s41598-025-00801-0

Huang, Y.-M., & Chiu, P.-S. (2015). The effectiveness of a meaningful learning-based evaluation model for context-aware mobile learning. British Journal of Educational Technology, 46(2), 437–447. https://doi.org/10.1111/bjet.12147

Johnson-Glenberg, M. C. (2008). Fragile X syndrome: Neural network models of sequencing and memory. Cognitive Systems Research. 9 (4): 274–292. doi:10.1016/j.cogsys.2008.02.002.

Kestin, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458. https://doi.org/10.1038/s41598-025-97652-6

O'Brien, D. (2021). A guide for incorporating e-teaching of physics in a post-COVID world. American Journal of Physics, 89(4), 403–412. https://doi.org/10.1119/10.0002437

O'Brien, D., Vieyra, R., Cortés, C., Johnson-Glenberg, M., & Megowan-Romanowicz, C. (2023). Evaluating learning of motion graphs with a LiDAR-based smartphone application. arXiv:2301.10334. https://doi.org/10.48550/arXiv.2301.10334

Paul, R. & Elder, L. (2007). The Art of Socratic Questioning. Dillon Beach, CA: The Foundation for Critical Thinking.

Rodríguez, E. Q. (2025). The body is the key: A formal architecture for embodied grounding in artificial general intelligence (Introducing the virtual embodiment module. https://doi.org/10.13140/RG.2.2.11025.70244

Sharples, M., & Pea, R. (2014). Mobile learning. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences (2nd ed., pp. 501–521). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526

Vieyra Software. (2025). Physics Toolbox. https://www.vieyrasoftware.net/

Vieyra, R., Megowan-Romanowicz, C., O'Brien, D., Cortés, C., & Johnson-Glenberg, M. (2023). Harnessing the digital science education revolution: Smartphone sensors as teaching tools. In Handbook of Research on Digital Teaching and Learning (Chapter 8). IGI Global. https://doi.org/10.4018/978-1-6684-5585-2.ch008

Vieyra, R. E., Megowan-Romanowicz, C., Johnson-Glenberg, M. C., O'Brien, D., & Cortés, C. V. (2024). Making motion meaningful: Mapping body movements onto graphs. The Science Teacher, 91(6), 57–64. https://doi.org/10.1080/00368555.2024.2404956