RAPID: DRL AI: Unlocking the Potential of Generative AI for Equity and Access in Robotics Education

STEM (Science, Technology, Engineering, and Mathematics) education underpins the efficacy of America's future workforce, and in turn its prosperity and national security. Equity in education ensures that the entirety of the nation's diverse population can both contribute and benefit. Yet while most technological innovations come at the expense of equity, amplifying the output of only the most-expert, emerging evidence suggests that Generative Artificial Intelligence (AI) may be the rare technological unicum which inseparably advances both productivity *and* equity.

Unlocking the Potential of Generative AI for Equity and Access in Robotics Education (UP-GEARED) investigated this potential in the context of educational robotics – a highly visible and widely adopted STEM setting that facilitates rapid scaling and impact. Specifically, it examines the educational equity potential of **task-oriented generative AI assistants** that help directly on skilled, creative productivity tasks such as engineering design.

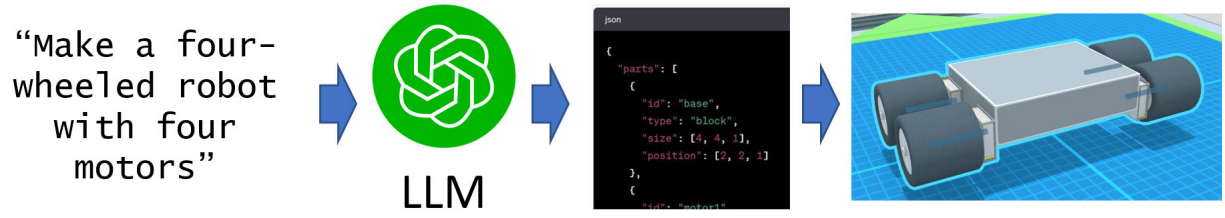
We hypothesized that not only would such an assistant increase student design success overall, but this assistance would silently bridge prior "nuts and bolts" knowledge gaps between more- and less-experienced individuals, allowing disadvantaged learners to successfully engage in high-level STEM activities. Eventually, such a system would help to narrow long-standing equity gaps in performance, learning, and motivational outcomes.

1.1 Effort Overview

UP-GEARED consisted of two primary activities: (1) Developing our proof-of-concept AI Robot Design Assistant into a classroom-ready prototype; and (2) Conducting a classroom study to test for the hypothesized benefits of AI Assistant Use in attenuating prior experience gaps. Our primary findings are about the design of task-embedded generative AI systems for engineering education, and their effects on diverse student performance on a virtual robot design task.

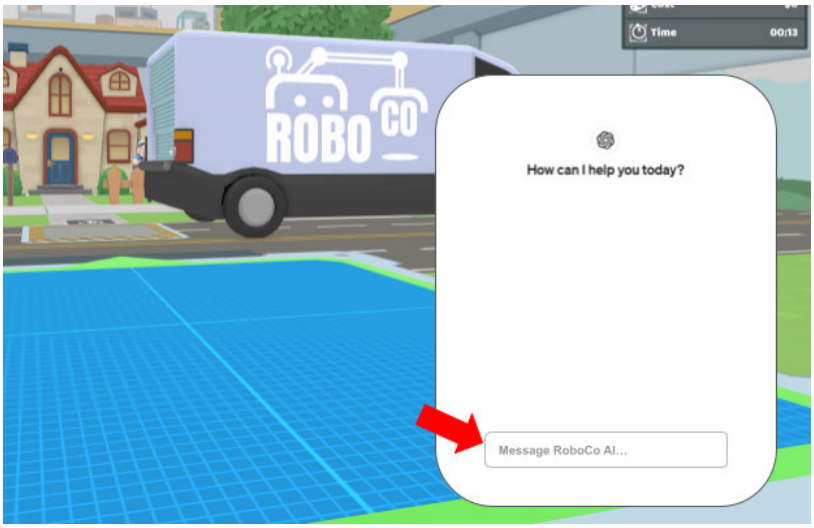
1.2 Developing an AI Robot Design Assistant

We developed and tested an AI robot design assistant that generates playable robot designs and modifications in a virtual robot-building game called RoboCo, based on user prompts such as "build a four-wheel driving base" and "add an arm". We developed a scripting language that describes in-game robots as labeled assemblies of parts. We then used an off-the-shelf Large Language Model (LLM) – ChatGPT 4o in the final prototype – to generate code in that descriptive format, creating functional robots in the game.



1.2.1 Understanding Likely Queries

The user interface was informed by an online social media survey we conducted in which we showed viewers a mockup of an LLM prompt in a RoboCo level and asked them what they would type in the box below.



We collected *n*=112 responses and examined them to derive *types of queries* that the assistant would need to be able to handle, and the *types of responses* that would be needed. We arrived at six distinct system behaviors based on the type of query detected: requesting clarification, giving engineering process advice, describing a robot design (either the player's or one from the system's library of known-good designs), suggesting applicable components or constructs from the library, generating a construct from a user description, or modifying a construct as described by the user.

1.2.2 Final System Architecture

The in-game AI feature/agent was named Tavix and integrated into a special build of the RoboCo game called RoboCo AI. Tavix is an OpenAI gpt-4o LLM agent that utilizes a decision tree to first characterize each incoming user query, then route it to an appropriately tailored response chain. Context including current task goals, current robot, list of available robot parts, and RoboCo simulation knowledge are injected as context into relevant system prompts in the prompt chain. A memory of agent-generated robot assemblies and the chat history are preserved to allow iteration within chat sessions.

This approach allows the system to handle a very broad range of queries by directing them to specialized response chains, while retaining context across both the simulator and chat spaces. In this way, Tavix can handle complex chained interactions – suggesting blueprints and parts to jump start task-specific robot design, altering the user's robot design to demonstrate high level engineering techniques (optimization, mass balancing, gear ratios, etc.), and discussing the resulting design with the user.

Tavix is built on widely available, reusable infrastructure. The Tavix processing pipeline was built in Python using LangChain; the user interface was developed with Streamlit. The chat interface and backend are served in AWS Elastic Container Service. The prompt chain relies on JSON structured output via OpenAI tool calling in all but the final step in the chain. This allows for type validation of each step and ensures that the final agent-generated assemblies can be rendered in RoboCo.

1.3 The Effects of the Robot Design Assistant on Student Performance

Previous studies have shown Generative AI assistance on core work tasks narrows gaps in performance between new and experienced workers, and improves retention of those newer workers (Brynjolfsson, Li, & Raymond, 2023). We wanted to know whether analogous performance gap-closing and motivational effects would emerge in a classroom context.

1.3.1 Classroom study design

We conducted a crossover-design study with two sections of a high school CAD class (*n*=31) in late 2024. Both sections played through four increasingly difficult RoboCo robot design challenges. One class section was given access to Tavix only in the first half of the assignment, while the other section received it only in the second half.

- A1 vs. B1: Does having AI help?
- A1 x B1 x Experience: Does AI narrow performance gaps?

At the beginning, at the switch-over point, and at the end of the lesson, learners filled out a short survey about their confidence in their ability in STEM and Robotics (self-efficacy). The first survey also asked about prior robotics experience and optional demographic information. The midpoint and post-surveys asked whether they had noticed an AI helper and whether it had been helpful on that day. As students played through the RoboCo challenges, the software recorded the number of attempts and amount of time each player took to complete each level, the number of queries they submitted to Tavix, and the content of each query.

1.3.2 Classroom Pilot Study Results

66% of students who completed the questionnaires reported noticing Tavix at some point during the activity. 68% of students who noticed Tavix said they used it. This roughly matches our logs, which indicate 48% of all students (12 of 25) actually used Tavix during an activity. Of those who used Tavix in the first half, 82% rated it as helpful. 62% rated it as helpful in the second half, perhaps owing to the higher difficulty of later challenges.

We then compared time taken by Tavix users vs. non-users to complete the flagship "Bistro" challenge – the first significantly challenging in-game level. Tavix was available to 52% of the students on this challenge (13 of 25), and used by 24% -- a **46% uptake** rate. Those who used Tavix had significantly worse performance – 50% of Tavix users completed the activity, versus 84% of non-users; Tavix users also spent 85% longer on the challenge (mean: 45 minutes vs. 24.3 minutes), and logged an average of 32 attempts in that time compared to 14.6 attempts (2.19x as many). This suggests that Tavix use is **negatively associated** with activity completion. However, the **directionality is unclear** – it could be that Tavix use is inefficient, or that students came to Tavix for help when they were not succeeding.

This phenomenon is clearly in need of further unpacking, particularly since the classroom teacher report indicated that students who used the AI assistant on that particular challenge were making *better* progress and experiencing less frustration than students who did not. There was no clear association between prior robotics experience level and Tavix use, though we can observe that the heaviest Tavix users (5 queries submitted per user) eventually ended up completing the activity, while three of the four

light users (1-2 queries each) did not. Is it the case that users who "lean into" the AI assistance benefit while those who only briefly try it to get unstuck do not? Given the small sample size of *n*=6 active Tavix users on that challenge, we can only speculate until further studies are conducted.

1.4 FUTURE WORK

One of the hallmarks of exploratory research is it leads to new questions that could not have been asked at the start. For instance, in addition to the cases discussed above, at least two students disregarded study instructions and used Tavix the entire time – one of these students had high prior experience and the other had none. Do different groups of learners engage with Tavix (perhaps even in defiance of expectations) for different reasons? Further analysis is continuing on pilot study data. We hope to learn more about this and other potential patterns by narratively reconstructing a small number of cases of Tavix use through detailed logs. Do learners mostly ask Tavix for help up front or after getting stuck? Do they ask for advice at the beginning and designs at the end, or bounce back and forth?

Additionally, while our generative AI assistant "Tavix" now represents a stable, reusable, and commercially deployable agent for STEM education – as well as a considerable improvement in capability over its earlier iterations – there is still room for growth in the quality and sophistication of its output. In particular, Tavix is quite effective at providing learners with relevant prefabs (known-good robot part templates from our library), but only vaguely proficient at from-scratch generative tasks like creating a functioning arm assembly, and prone to error when asked to edit an existing design. One likely area for improvement is to use more advanced forms of knowledge incorporation such as Retrieval Augmented Generation (RAG) to allow the system to generate outputs derived from a larger known-good robot library, and to extrapolate from relevant designs rather than selecting among them.

Notably, though, the exact mode of failure at robot generation is typically physical in nature – for instance, parts joined or overlapping in impossible ways. This effect persisted even after numerous algorithmic attempts to reject and regenerate such output. We believe this points to a specific capability gap in current LLMs' physio-spatial understanding – one that makes sense, since off-the-shelf LLMs such as ChatGPT 4o and Llama 3 are trained on text and images, not physical models or interactions beyond those implicit in pictures. There are efforts in the robotics space to remedy the lack of training data in physical modalities, for instance Nvidia's efforts to use PhysX-accelerated simulations to scale up reinforcement learning in this area. Tavix could not only benefit from the incorporation of these technologies, but serve as a testbed for their generalizability.

Finally, the classroom studies should be continued and expanded. This first-round study was limited to comparisons of student performance based on task completion – the true educational value of the system would be impact on learning.

This material is based upon work supported by the National Science Foundation under Award No. 2341190. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1.5 REFERENCES:

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work (No. w31161). National Bureau of Economic Research.*