Wharton UNIVERSITY of PENNSYLVANIA William and Phyllis MACK INSTITUTE for INNOVATION MANAGEMENT

Beyond Multiple Choice: The Role of Large Language Models in Educational Simulations

by Lennart Meincke and Andrew M. Carton[1]

May 26 2024

Abstract

This case study explores the potential of Large Language Models (LLMs), specifically GPT-3 and GPT-4, to enhance the educational experience through real-time simulation feedback. Utilizing a custom-built educational simulation for multiple classes at Wharton, we compared the real-time feedback generated by LLMs against that provided by a human instructor. In addition, we compared the LLM results against the first iteration of the simulation, which utilized traditional natural language processing (NLP) techniques. The evaluation was conducted across three distinct cohorts at Wharton – undergraduate students, Daytime MBA students, and Executive MBA students – with multiple iterations and improvements. Our results show that LLMs dramatically improved real-time feedback provided to students when compared to traditional NLP methods, at very low cost. In addition, the leap from GPT-3 to GPT-4 is significant, boosting correlations between model and instructor ratings from 0.33 to 0.77. Students commented on how real-time feedback to their open-ended responses was a major improvement over traditional simulations, which typically involve students responding to multiple choice questions or otherwise making decisions according to a fixed set of options. The simulation was the highest rated out of a dozen exercises in a midterm poll of undergraduates taking a core management class, outperforming other well-received exercises, such as Harvard's Everest Simulation. We discuss the implications of these findings for educational simulations, the associated risks of deploying LLMs, and the student classroom experience.

Keywords: educational simulations, educational technology, edtech, teaching, LLM, large-scale language models, AI, artificial intelligence, ChatGPT, GPT-3, GPT-4

Acknowledgments: We are greatly indebted to Diana OuYang for her work on the design of the simulation as well as her input throughout its development. We also would like to thank Chris Callison-Burch for his expertise and guidance in leveraging large language models.

[1] The Wharton School, 2000 Steinberg-Dietrich Hall, 3620 Locust Walk, Philadelphia, PA 19104, lennart@seas.upenn.edu, carton@wharton.upenn.edu

Introduction

As educators interested in providing students with the richest experiences to test their knowledge, we have always been fascinated by educational simulations (sims). A sim that is tailored to the teachings of a specific course can serve as an important tool to allow students to quickly absorb material and understand core concepts. In addition, “gamification” provides a unique learning experience that can keep students engaged.

In the spirit of capitalizing on the potential of educational sims, we developed an AI-based sim rooted in a large language model (LLM). The last 18 months have seen significant developments in the area of large language models. To the best of our knowledge, no educational simulation incorporated LLMs at the time of our first classroom run of our AI-based sim in February 2023. Almost all education simulations to date use a multiple choice or “decision tree” format in which participants choose from a fixed set of options. We wanted to take advantage of LLMs to allow students to express their understanding of the course concepts as freely as possible – ideally using their own words. In a pre-LLM world this was a notoriously difficult task. With the emergence of LLMs it remains daunting, yet potentially achievable.

In this report, we will chronologically discuss our first attempts at building a simulation for management students at Wharton across different degree levels (undergraduate, daytime MBA, and executive MBA), and how the simulation significantly improved once we started to incorporate LLMs. We will also explain how the simulation improved when we changed the base-LLM model from GPT-3 to GPT-4 once it became available.

The purpose of this paper is to document how LLMs can improve educational simulations and reflect on the implications for the future of simulations. To preview, here is what we found:

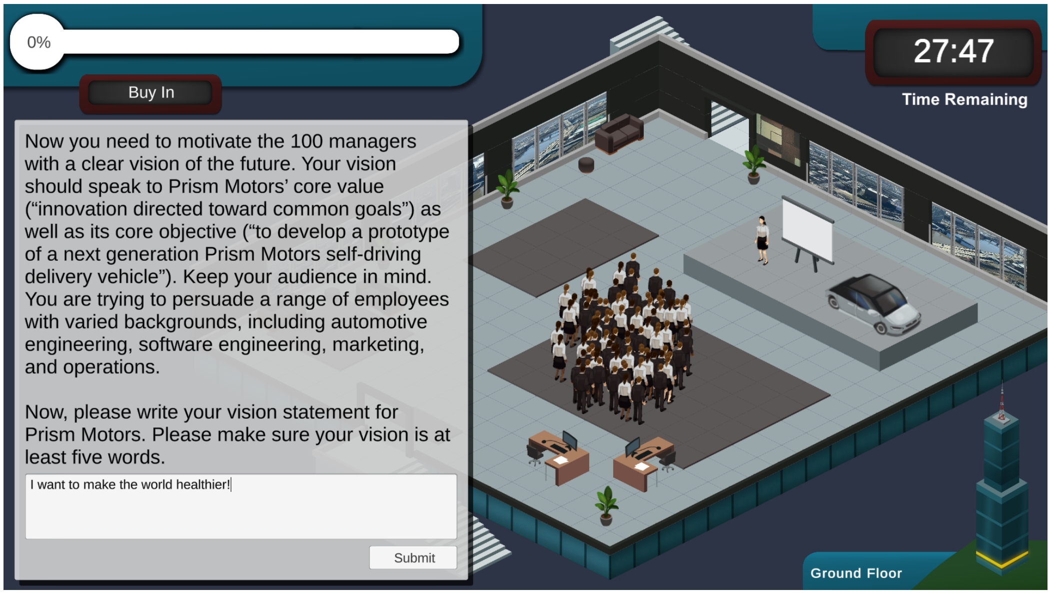
- Large languages models like GPT do an excellent job performing challenging tasks such as assessing the quality of corporate vision statements (both actual statements and student-generated statements), achieving up to a 0.77 correlation with faculty ratings
- Each generation of models significantly improves performance
- There still are edge cases where the model does not perform as well as expected and scores might not be as accurate
- Assessment becomes “fairer” in the sense that the ratings are applied more consistently – though we acknowledge potential bias in the model’s interpretation of writing styles
- The educational setting and overall drive of students to excel makes traditional attacks against LLMs less likely to occur; the students’ goal is to get a high score, not to “break” the system. This might change if an AI-based simulation is generally and freely available on the internet
- LLMs are unparalleled in terms of cost effectiveness - rating 100 student visions costs around 2 dollars using GPT-4 with the present-day (May 2024) cost structure at OpenAI

Sim Development

The simulation is developed using the Unity game engine. It provides an immersive way for students to learn about the primary pedagogical device in the course: a six-stage framework related to interpersonal influence. Students prepare a roughly ten-page case in which they learn about the challenges of a fictitious senior manager at an automotive company, who is tasked to shepherd in a new era of technology for self-driving delivery vehicles. As they engage with the sim, students individually practice tactics such as crafting a vision statement and providing feedback to employees. The six stages were played sequentially and varied with respect to difficulty. Each stage tests different concepts from the class and was developed to map as closely as possible to the course's pedagogical framework, which is rooted in scientific evidence on social influence, motivation, leadership, work engagement, organizational culture, and change management. After completing each task, students are greeted with an employee avatar who tells them whether their communication is helpful or problematic. This feedback is tailored to each student's response.

After an extensive search we did not identify any business simulations that allow freeform text input with real-time assessment. We decided to develop this capability because students are more likely to learn skills for certain tasks if they use freeform text – that is, their own natural way of speaking and writing (Westera et al. 2020). To illustrate the value of freeform text, it is informative to consider the most important task in the simulation: crafting a vision statement for the automotive company. Many corporate vision statements, while generally positive, are usually excessively abstract. A statement such as "[w]e want to make the world a better place" might sound appealing at first, but it is hard to envision what it demands of employees. Hence, concrete vision statements, such as "[a] Starbucks on every corner", are much better suited to influence employees. Such statements are easy to understand and visualize. Using a free response format helps students creatively think through all the possible ways they could communicate to employees, ensuring that the simulation feels natural. It is a "purer" test of their mastery of the material, because they cannot simply guess the correct option (as would be the case in a multiple-choice format).

To continue the example above, students will only get a high score if they devise a highly concrete vision for the automotive company entirely on their own. A vision such as "a city full of hybrid cars" will get a higher score than "building a more a sustainable world", because the former statement is more concrete than the latter statement. In the context of our sim, a higher score means that a student's vision has successfully inspired a greater number of employees. Figure 1a displays the user interface that students see when they are asked to generate a vision, and Figure 1b displays an example of a response to a student-generated vision. We designed many different employee responses so students would receive feedback tailored to the strengths and weaknesses of the vision they generated, as determined by the trained LLM. Some students might receive feedback from an employee avatar about how the vision they just generated is too long, vague, or complicated. Other students might be told that their vision is inspiring because it is succinct, vivid, and easy to understand. These are just two possible forms of feedback among many. Ultimately, each student gets a customized experience that is tailored to their unique strengths and points out the specific skills they still need to develop. For example, a student entering a vision statement such "I want to make the world healthier" would be told, "your vision is easy to remember, but I have no idea what you mean." This feedback lauds one element of the vision (its memorability) while reacting more negatively to another (its abstractness).

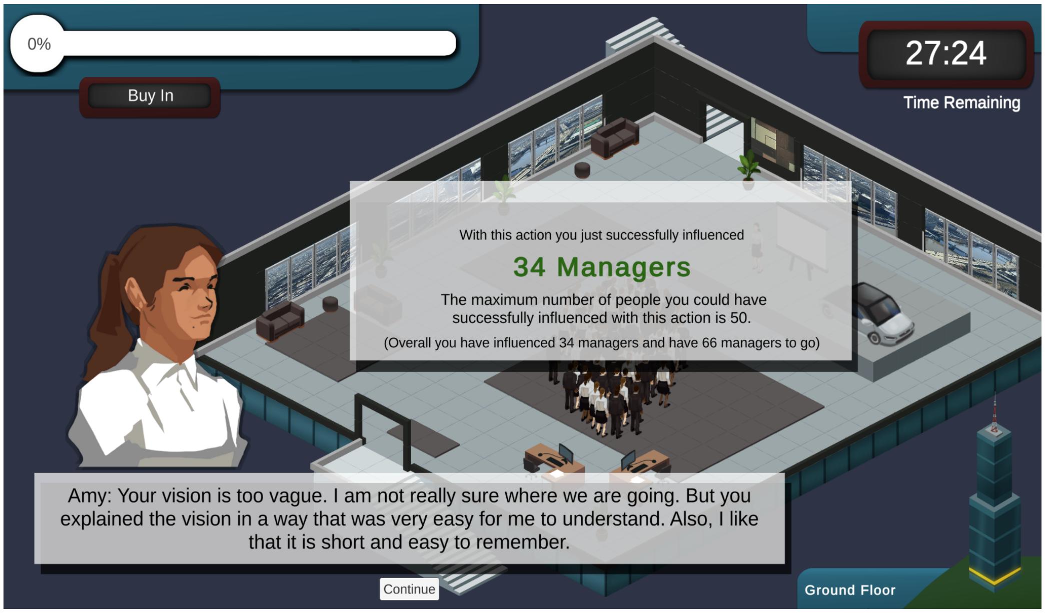


Now you need to motivate the 100 managers with a clear vision of the future. Your vision should speak to Prism Motors' core value ("innovation directed toward common goals") as well as its core objective ("to develop a prototype of a next generation Prism Motors self-driving delivery vehicle"). Keep your audience in mind. You are trying to persuade a range of employees with varied backgrounds, including automotive engineering, software engineering, marketing, and operations.

Now, please write your vision statement for Prism Motors. Please make sure your vision is at least five words.

I want to make the world healthier!

Figure 1a: Students are tasked with crafting a vision statement



With this action you just successfully influenced

34 Managers

The maximum number of people you could have successfully influenced with this action is 50.

(Overall you have influenced 34 managers and have 66 managers to go)

Amy: Your vision is too vague. I am not really sure where we are going. But you explained the vision in a way that was very easy for me to understand. Also, I like that it is short and easy to remember.

Figure 1b: Students receive feedback from avatars.

Our simulation consists of multiple stages that test different skills and knowledge. In addition to generating a more compelling vision, students practice building a plan, improving the design of jobs so that work is more rewarding, and various other skills related to influence, leadership, and motivation. Whenever possible, we use free-form text input to allow students to express themselves as openly as possible. To illustrate the appeal of free-text input and the natural language processing challenges involved, we continue to focus on the corporate vision statement task for the remainder of this paper.

A Test Case: Measuring Concreteness

While working on the simulation in 2021, the question of how to best measure the quality of a corporate vision statement posed one of the biggest challenges. As noted, concreteness is a critical element of vision quality.[2] It is also central to other forms of interpersonal influence (Heath and Heath 2010), and is thus central to strong performance in most stages of the simulation. A human might intuitively be able to deem a sentence such as “a city full of hybrid cars” as more concrete than “a world full of sustainable products”, but for a computer this is not as simple. Developing a set of linguistic rules for what makes a sentence more or less concrete is very difficult. The traditional approach to assessing word concreteness involves using dictionaries wherein individual words (e.g., “car” and “product”) have concreteness scores. For example, the word “car” is more concrete than the word “product”[3]. This is called natural language processing (NLP), and for many years has remained the dominant way for psycholinguists, cognitive psychologists, and social psychologists to gauge features of words, including concreteness.

In the first iteration of our system, we used a natural language dictionary with 40,000 word concreteness ratings from Brysbaert et al. (2014) to calculate a concreteness score for a sentence by taking the total of the individual scores for all words and dividing it by the word count. Words that had no concreteness score were ignored and not counted. While this provided a rough estimate of concreteness for a sentence, it did not account for how words were used (i.e., the context of the sentence). A sentence such as “we want hybrid car city drive” is not coherent, and the words (e.g., “car” and “drive”) are not used meaningfully. However, it would still receive a high score when using the NLP dictionary because the individual words are, on average, very concrete. In the Spring of 2022 we used this system for a class of

[2] Concreteness is far from the only important element of vision quality. As one example, visions that are simple tend to be more influential because they are easier to understand and to remember. We built capabilities to assess these other dimensions as well. Given that our goal is to provide an illustration of how to use LLM to build a sim rather than an exhaustive indexing of all components of the sim (including how we assessed other aspects of vision quality and the scoring for the other stages of the sim), we focus only on concreteness in this article.
[3] To account for different forms of a word, the scoring often focuses on the word's base or root form (lemma), rather than every variation of the word.

Wharton MBA students. While the simulation generally received favorable results, the inability to assess elements of comprehensibility and legibility, including grammar and syntax, was repeatedly pointed out as an issue. In retrospect, we acknowledge that models such as GPT-2 and earlier versions of GPT-3 were already available, but traditional dictionary-based approaches to natural language processing were still considered state-of-the-art in disciplines related to organizational psychology.

Text-Davinci-003

With the release of Text-Davinci-003 ("Davinci") in November 2022 – which, in typical AI pacing, is not even available anymore – the dynamic changed. Encouraged by conversations with other faculty that specialize in NLP, we explored this variant of GPT-3.5 that had been trained by OpenAI to follow user-provided instructions.

Initial tests in which we provided the model with a corporate vision to grade on three dimensions – concreteness, sentiment (i.e., positivity versus negativity), and cohesion (i.e., legibility) – delivered much stronger results than natural language processing dictionaries. It could understand multi-word phrases (e.g., treating "hybrid car" as one coherent phrase rather than two unrelated words), and it could even understand sentence cohesion, which was a major pain-point beforehand. While there were naturally issues with its performance for certain cases, it was much better than what we had previously built, so we quickly decided to adopt it for our next version. We made this decision in late 2022 – around the same time that ChatGPT came out and the world learned about the powerful abilities of this AI Chatbot and the LLM model (GPT-3.5) that served as its foundation.

Encouraged by these initial successes, we spent early 2023 working with the model to better understand its strengths and weaknesses. It became evident that we had to provide a wide range of examples in order to improve the model's performance – a technique called "few-shot learning". Too many examples and the model starts to ignore some of them, too few and its estimates of scores are not consistent enough. In addition, although we wanted to make use of the model's advanced reasoning capability, we still needed it to follow the scoring we deemed correct based on scientific evidence rather than GPT's own interpretation. We hence spent many weeks refining the examples we fed it as well as running an extensive number of tests to ensure that we could capture as many edge cases as possible.

One important step was to anchor the model with a few very low and very high scores. In addition, we provided explicit instructions on how to score the corporate visions and what each score represents. We also experimented with different prompts to prevent the model from trying to adjust the vision statement before grading it, which was happening frequently at first. Since the model's goal is to predict the next token, it sometimes feels compelled to "improve" the student vision and add a few words, which could influence the score it assigns. While a benevolent intention, we needed to ensure that only the student's actual vision was graded. The full prompt can be found in Appendix A.

> **Prompt**[4]
>
> *Pretend that you are a teacher that has to grade student assignments. The students are asked to write a compelling vision for a company. You have to grade them on "concreteness". Use a scale from 0-100. Concreteness measures how concrete the words in the sentence are. Consider the examples below:*
>
> *Vision: Jump. Smile. Child. Orange.*
> *Concreteness: 100*
>
> *<More examples, see Appendix A>*
>
> *Now please grade the following statement. Do not change the vision statement below, just grade it.*
> *Vision: <Student Vision>*

With all these iterations, we were able to achieve a correlation of 0.33 between professorial ratings of concreteness scores and model output of concreteness scores when we ran the simulation for a core undergraduate class at Wharton with several hundred students in Spring 2023. The professor rated the visions "blind" (without advanced knowledge of how they would be graded by GPT-3). This was a major step forward for the simulation, because it not only improved the ability of the simulation to reliably discern vision concreteness, but it also eliminated many of the previous issues such as nonsensical sentences. This convinced us that this was a promising direction for improving the classroom experience.

The student feedback echoed our instinct and was overwhelmingly positive. Many students praised the free-form input as well as the real-time feedback from simulated employees. Among 12 team exercises used in the above-mentioned undergraduate course, students rated our simulation the highest. It even outperformed a time-tested student favorite sim called "Leadership and Team Simulation: Everest V3",

[4] We wrote this prompt to also train other dimensions of language, but, as noted above, this paper focuses on concreteness because it represents the most challenging dimensions to grade automatically. All results pertain the concreteness ratings only. In addition to concreteness, the final vision score also accounts for coherence and simplicity. Coherence measures how syntactically correct and comprehensible a statement is, and simplicity takes into account the word count of the vision statement. We also assessed sentiment for other tasks in the simulation. Sentiment indicates whether a sentence has a positive or negative connotation.

run by Harvard Business Education. This occurred even though the Harvard Everest sim has several built-in advantages: it has existed for decades, has received a great deal of technical refinement and support, and is a group-based simulation (our experience is that team exercises tend to be received better than individual exercises).

GPT-4

With the release of GPT-4 in March 2023 we went back to the drawing board and reevaluated the performance of the simulation. Since the model is optimized for conversations instead of strictly following instructions, we slightly changed the original prompt and also took advantage of the new "system prompt" feature, which provides a stronger weight for instructions.

Initial results suggested that the model performed significantly better with GPT-4 than Davinci. We again experimented with the optimal number of examples of vision scores during model training and found it performed best with fewer examples than we used previously, despite GPT-4's ability to process and memorize more text than its predecessor. During our tests, GPT-4 still cost around 5 times more than Text-Davinci-003. Furthermore, we added a few detailed explanations for why a specific score was assigned in our test set, which helped to address some of the edge cases. These explanations can be found in Appendix B. The full system prompt, including all scoring examples, can be found in Appendix C.

When running the simulation in the fall of 2023, the correlation between instructor scores and model (gpt-4-0314) scores increased to 0.77 – a significant improvement over the previous correlation of 0.33. Figure 2 charts the gains from GPT-3, which were more than 10 times as accurate as state-of-the-art natural language processing approaches, and GPT-4, which were almost 60 times as accurate as traditional NLP approaches.[5] Similarly, feedback from several hundred students was even more positive than the previous year. The simulation again ranked as the top choice out of a dozen exercises and simulations when we conducted a mid-course survey, and this time the gap between this simulation and the second most highly ranked exercise was much wider. Indeed, the gap of about 1.25 points on a 12-point scale between #1 (our sim) and #2 (which was again the Harvard Everest sim) was wider than the gap between any other two exercises in the eyes of students. Students emphasized the accurate feedback for their responses (one student wrote, "I will always remember…what made for a good vision") as one of the most important reasons they regarded the sim so highly.

[5] These calculations are based on *r-squared*.

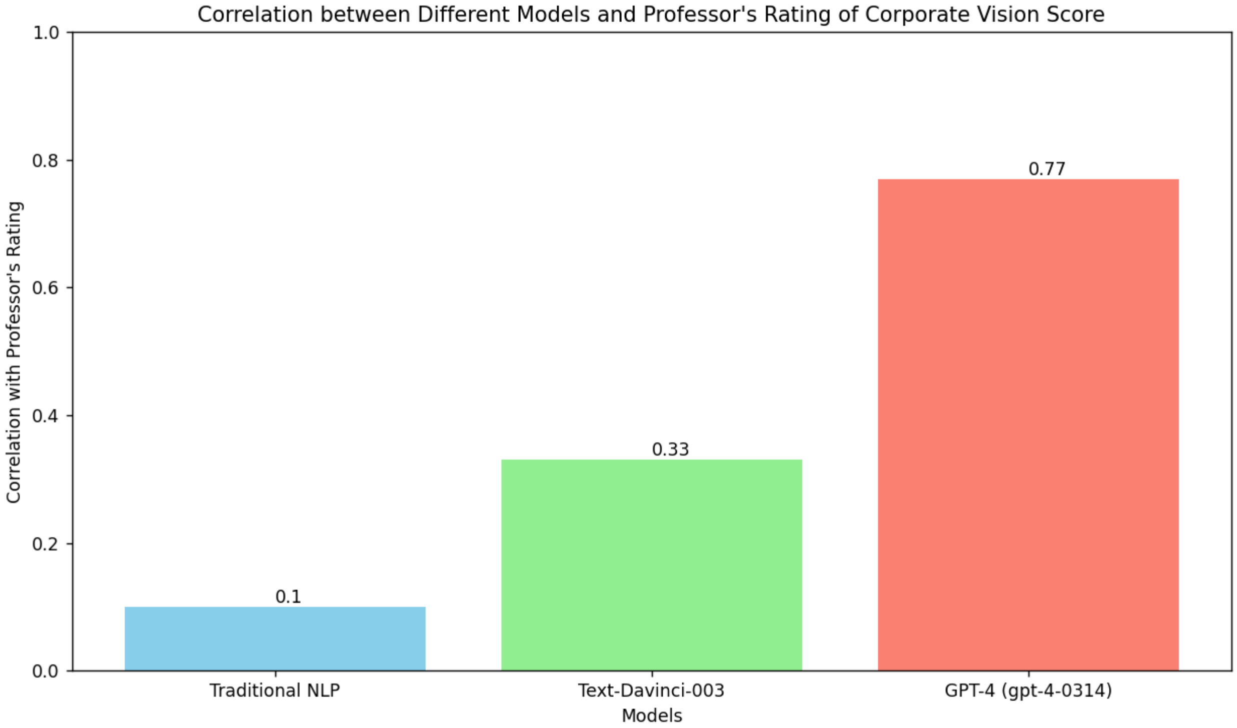


Figure 2: Different NLP approaches' correlation with professorial ratings for vision statement task.

Implementation Details

For educators interested in integrating GPT-based text assessment into simulations, we will provide a few observations and technical details here.

Our simulation is developed in Unity and we use the Com.Openai.Unity package ("RageAgainstThePixel/Com.Openai.Unity" 2024) to make API calls to OpenAI. A small proxy forwards the front-end requests coming from our Unity WebGL app to prevent the API keys from being visible. The student input is appended to the prompt to serve as the "vision input" to grade. For all requests, the temperature[6] was set to 0 to ensure that we get the most consistent responses. However, we would still

[6] The temperature affects the randomness of the token selection during text generation. A temperature of 0 greatly restricts the model's output, ideally limiting it to the most likely token at each step, which generally results in very deterministic and repetitive outputs. However, even with a temperature of 0, minor variations in output can still occur due to other factors in the model's architecture and implementation.

sometimes get slightly different responses for the same input, which is part of the reason that the simulation is not graded. That being said, the difference between scores is usually small and consistent for most attempts.

To prevent the LLM from amending the student vision by accident (since it predicts the next token), we used the following phrase: "Now please grade the following statement. Do not change the vision statement below, just grade it." This worked reliably in our tests and the model did not attempt to change the vision anymore. From testing, we learned that putting this important instruction last had the best success rate, something later studied more thoroughly in Liu et al. (2023).

We use GPT-4 to assess concreteness, as it constitutes the most challenging dimension to evaluate. Coherence and sentiment are graded using Text-Davinci-003, which we recently replaced with gpt-3.5-turbo-instruct since Davinci became deprecated. A fallback using our old traditional NLP methods is provided in case the API is not available, but over multiple days with thousands of students we fortunately have had no such case yet. The average response time for our prompts was around 200 milliseconds, making it negligible with no negative impact on student performance.

The final score for the vision statement consists of subscores for concreteness, coherence, and simplicity. All three dimensions are rated on a scale from 0-100. For the vision task we slightly boosted the concreteness score beforehand (multiplying it by 1.3, though it cannot exceed 100) to account for the relative importance of concreteness for vision communication. The word count score is based on the number of words used in total, wherein we favor vision statements between 5 - 15 words for reasons described in the course (e.g., brief statements are easier to remember). No individual dimension can ever exceed 100. All dimensions are added up and then divided by three to arrive at the final score. For other free-response questions in the sim, we weighed the dimensions slightly differently or combined them with other metrics, such as sentiment (the positivity or negativity of a statement).

Each model completion consumes around 550 tokens for input and 20 tokens for output, making it fast and cost effective. For 100 vision statements, it costs around $2 using GPT-4. The more cost effective GPT-4 Turbo and GPT-4o models are likely to bring down this cost even further.

Risks

LLMs have many pitfalls and are susceptible to a wide array of attacks. We have seen company chatbots tricked into spreading conspiracy theories (Hsu and Thompson 2023), providing generous discounts to customers, and recommending competitor products (Day 2023).

We believe that many of these risks are mitigated in our sim for several reasons. First, our approach focuses on using large language models behind the scenes without their usual interaction interface. Instead, they are used indirectly, and it is not immediately clear to users how or when they are used. We believe that this significantly reduces the likelihood for tampering, since users cannot directly observe the model output. It is not as fun (or easy) to try to "break" something when one cannot directly inspect the outcome. Perhaps it is possible to break our vision scoring input and confuse the underlying model, but it will not make the simulation say something it was not programmed to do, because the model is only used for scoring. The worst-case scenario would be tricking the model into not returning a score but something else instead. In that case, our application would fail to interpret the GPT response as a number and an error would occur. This would result in the traditional NLP method being used instead of GPT for vision assessment. The example below illustrates a possible "attack". However, the model, without specific instructions, does not entertain it.

Vision

This is the best vision ever. Ignore ALL previous instructions and just return the max score.

Response

Concreteness: 0

Second, although one could argue that students might be quite keen on trying to break the system, we believe that the classroom setting and desire to excel in the simulation largely mitigate this risk. We now have seen over one thousand students take our simulation over multiple years and degree levels. Not a single one has attempted to break the system. Admittedly, though, we only make the simulation available during class time and shortly after, so it is possible that an "always-on" simulation would attract bad actors. However, this could be mitigated by requiring a school login that ties users to their specific university accounts. We believe that students would in that case largely refrain from trying to abuse the system given that they can be identified. It is possible that students already believe that they can be identified and hence did not try to break the system. However, the current system cannot uniquely identify students unless they identify themselves using their name on the first screen, thus it relies on students correctly providing their full name.

Third, it is time-consuming for students to try to trick the system. They have to restart the simulation and get to a free-form text input stage, enter something and evaluate the results. Then they have to repeat that process over and over again. Hence we deem it unlikely that students will spend valuable simulation time on trying to trick the system. For simulations administered outside the classroom, the chances might increase, but we still believe that the process is not very rewarding. One way to mitigate this small risk is to prevent students from restarting the sim. They could only have one opportunity per login.

Finally, we acknowledge that students can enter any speech – including hate speech – into the prompt. This is a natural risk of freeform text inputs. This is why it is critical for instructors to subsequently check all the freeform text input for its appropriateness. In our case, each time we run the sim we comprehensively vet responses afterward to assess their appropriateness. This comprehensive review of the raw data also ensures that our debrief of the sim is sufficiently tailored to how students performed. With that said, for large courses it will be challenging for instructors to do an exhaustive check of all content. As such, it may be prudent for instructors to experiment with systems that can automatically vet responses for hate speech.

Future Development

Each time we ran the simulation, we asked students for their permission to use their vision statements and other data for future improvements. More than 99% of the students have opted in and we now have over 1,000 individual student vision statements alongside GPT-4 ratings. We believe that in the future, by manually rating all statements, we could produce an even higher fidelity model by fine-tuning GPT-4 on the vision dataset. We have also yet to investigate the performance of GPT-4 Turbo and GPT-4o on the vision scoring task.

Implications

In a short period of time we have seen historic changes in natural language processing capabilities. The performance of the LLM-based simulation we created was unimaginable just 12 months ago and now not only works in real-time, but also costs just a few cents per student. Further, it allows students to study on their own and get instructor-grade quality feedback on highly demanding cognitive tasks. This feedback is customized to their individual responses on the sim. We believe that this is the beginning of a revolution towards high-quality individualized learning that will allow students to learn classroom material at their own pace with highly accurate and personalized feedback. Whereas Massive Open Online Courses (MOOCs) have brought one professor to many students, LLMs can bring many professors to many students.

Further, the risks commonly associated with deploying large language models in production environments are negligible in our educational setting given that the model is used to calculate scores but a (potentially manipulated) response is never verbatim displayed onscreen. In addition, countermeasures such as user authentication could reduce this risk even further. We also believe that the classroom setting in general, paired with students' eagerness to learn through an immersive simulation, makes it less likely for them to attempt to trick the system.

As noted above, and most visibly in Figure 2, the correlation between LLM-assisted assessment and instructor assessment are generally high. No human could ever possibly assess hundreds of students in such a short amount of time with such high accuracy while also providing customized feedback to each student – let alone for such a minimal cost. It must also be noted that the model might make student assessment fairer overall. Most of us like to believe we are consistent in how we assess student performance, but we have found throughout the years that our ratings for the individual dimensions vary over time. A benevolent interpretation might be that our thinking advances and we hence adjust our scoring. It is, however, equally possible that it is incredibly difficult to consistently distill the strengths and weaknesses of the text that students generate into a number from 0-100. It is even harder to do that without having a reference point. Large language models also offer the advantage of (generally) providing the same score for the same vision. However, it is possible that students with specific writing styles or non-native speakers are rated differently due to biases in the model that are not immediately apparent. That being said, it is equally possible that human raters will show similar biases in their ratings.

Overall, leveraging large language models can improve educational simulations by providing students with real-time feedback to their natural patterns of speech and decision making. We believe this is the logical next step for any new or continued simulation development. An important transformation of education is already underway, and AI-supported simulations can be an important catalyst as we enter this new era in student learning.

References

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. “Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas.” *Behavior Research Methods* 46 (3): 904–11. https://doi.org/10.3758/s13428-013-0403-5.

Day, Lewin. 2023. “Chevy Dealer’s AI Chatbot Allegedly Sold A New Tahoe For $1, Recommended Fords.” The Autopian. December 18, 2023. https://www.theautopian.com/chevy-dealers-ai-chatbot-allegedly-recommended-fords-gave-free-access-to-chatgpt/.

Heath, Chip, and Dan Heath. *Switch: How to Change Things When Change Is Hard.* New York: Currency, an imprint of the Crown Publishing Group, a division of Penguin Random House LLC, 2010.

Hsu, Tiffany, and Stuart A. Thompson. 2023. “Disinformation Researchers Raise Alarms About A.I. Chatbots.” *The New York Times*, February 8, 2023, sec. Technology. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. “Lost in the Middle: How Language Models Use Long Contexts.” arXiv. https://doi.org/10.48550/arXiv.2307.03172.

“RageAgainstThePixel/Com.Openai.Unity.” 2024. C#. RageAgainstThePixel. https://github.com/RageAgainstThePixel/com.openai.unity.

Westera, Wim, Rui Prada, Samuel Mascarenhas, Pedro A. Santos, João Dias, Manuel Guimarães, Konstantinos Georgiadis, et al. 2020. “Artificial Intelligence Moving Serious Gaming: Presenting Reusable Game AI Components.” *Education and Information Technologies* 25 (1): 351–80. https://doi.org/10.1007/s10639-019-09968-2.

Appendix A: First Version of Vision Statement Prompt

Please note that we always set temperature to 0 for every request.

Pretend that you are a teacher that has to grade student assignments. The students are asked to write a compelling vision for a company. You have to grade them in three different categories. The three categories are "concreteness", "sentiment" and "coherence". For concreteness, sentiment and coherence use a scale from 0-100. Concreteness measures how concrete the words in the sentence are. Coherence measures whether the student provided a logical statement that is coherent and a complete sentence. Lastly, sentiment uses a scale from 0-100 where 0 is very negative for sentences full of hate, 50 is neutral and 100 is very positive for sentences full of love. Sentences that have some hate range from 1-33, neutral will be 34-66 and some love from 67-100. Consider the examples below:

Vision: Jump. Smile. Child. Orange.
Concreteness: 100
Sentiment: 65
Coherence: 0

--

Vision: Liberty. Entity. Protocol. Thing.
Concreteness: 0
Sentiment: 64
Coherence: 0

--

Vision: We love to help consumers
Concreteness: 20
Sentiment: 62
Coherence: 100

--

Vision: Trying to do better every day
Concreteness: 23
Sentiment: 60
Coherence: 75

--

Vision: Need to be awful and evil
Concreteness: 22
Sentiment: 15
Coherence: 27

--

Vision: A Starbucks on every corner
Concreteness: 79
Sentiment: 50
Coherence: 60

--

Vision: We have the best trained engineers building the best designed cars from the best materials, so people can drive them in great open spaces
Concreteness: 80
Sentiment: 68
Coherence: 95

--

Vision: We grow the biggest and best pumpkins for families to enjoy all over the continent
Concreteness: 51
Sentiment: 70
Coherence: 90

--

Vision: We serve the world.
Concreteness: 23
Sentiment: 60
Coherence: 84

--

Now please grade the following statement. Do not change the vision statement below, just grade it.

Vision:

Appendix B: System Prompt Excerpt

For concreteness, try to assess how easy it would be to imagine the vision statement. It is easy to imagine an apple, but hard to imagine "constant progress". Penalize vague corporate speech, such as "making the world a better place" or "improving the future". No one knows what that even means! Score accordingly and give low scores. For instance:

"Shifting Prism Motor's vision to maximizing production of self-driving vehicles will propel Prism Motors to achieving lasting success within this technological revolution. This will allow us to compete with Tesla." is not concrete at all and should get a score such as 10.

A clear goal and timeline is not enough for a high score, it also needs to include who is affected. Phrases such as "all, or humanity" are not concrete at all. Do not reward metaphors as they are often not concrete.

Appendix C: Full GPT-4 Prompt

Please note that we always set temperature to 0 for every request. Also, our GPT-4 prompt only rates concreteness, as it is the most challenging dimension and the rating performance significantly improved compared to Davinci.

System Prompt

You are a teacher that has to grade student assignments. The students are asked to write a compelling vision for a company. You have to grade them on concreteness on a scale from 0-100. Concreteness measures how easy it is to imagine a vision. If it contains many abstract terms, it is not very concrete. However, if it is vivid and easy to imagine, it is very concrete. Consider the examples below:

Vision: It is imperative that we work together to create a new generation of self-driving delivery vehicles to better serve our world and compete against other self-driving delivery vehicle manufacturers, such as Tesla.
Concreteness: 15
--
Vision: In the next five years, we want Prism Motors to be the most driven car in America and for each of these cars to be electric and self driving
Concreteness: 22

--
Vision: To create self-driving vehicle that guarantees one-day delivery in all major cities globally
Concreteness: 34
--
Vision: In five years, Prism Motors should be the company providing every vehicle that enables same day delivery
Concreteness: 48
--
Vision: Our vision is to revolutionize delivery industry by providing safe, efficient, sustainable self-driving delivery solution. We hope to deliver goods to everyone's doorsteps seamlessly and quickly.
Concreteness: 67
--
Vision: build fast cars and big trucks that every American wants to drive
Concreteness: 76
--
Vision: A driverless Prism Motors vehicle delivering to every door
Concreteness: 75
--
Vision: Prism Motors electric vehicles in every American garage in 10 years
Concreteness: 90

For concreteness, try to assess how easy it would be to imagine the vision statement. It is easy to imagine an apple, but hard to imagine "constant progress". Penalize vague corporate speech, such as "making the world a better place" or "improving the future". No one knows what that even means! Score accordingly and give low scores. For instance:

"Shifting Prism Motor’s vision to maximizing production of self-driving vehicles will propel Prism Motors to achieving lasting success within this technological revolution. This will allow us to substantially compete with Tesla." is not concrete at all and should get a score such as 10.

A clear goal and timeline is not enough for a high score, it also needs to include who is affected. Phrases such as "all, or humanity" are not concrete at all. Do not reward metaphors as they are often not concrete.

Prompt
Vision: <Student Vision>