ARDOR AND DILIGENCE

A Study on Rhetoric and Generative AI in Physics Courses
492R Capstone Project Report

Kayson Reardon

Brigham Young University | December 2024

Ardor and Diligence

A Study on Rhetoric and Generative AI in Physics Courses

Kayson Reardon

A senior capstone submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Bachelor of Science

Dr. Adam Bennion, Advisor

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Using an introductory physics course with N = 28 volunteer students, students were provided with an experimental AI chatbot and an introductory physics text sample written by the student researcher. The impact upon student's academic performance was analyzed and individual perception of the sciences was surveyed before and after introducing the materials. Examination of test scores revealed no statistically significant results from the experimental group as a whole. Self-reporting from students exposed to a traditional textbook and AI chatbot revealed twice as much engagement with the chatbot than the textbook. Addition reporting from students suggests that opinions on AI utility are broad and divisive. Finally, it is evident from this study that providing materials and resources intentionally designed to incorporate solid pedagogical backing is insufficient preparation to significantly impact a student's academic achievement in a class.

Keywords: education, generative AI, physics, terminology, rhetoric

MOTIVATION

Through personal experience as a student pursuing a degree in Physics, I know firsthand that some of the most enlightening instruction has come from collaboration with my peers. While there is much to say in favor of collaborative work, I believe that a large portion of the shared success has come from shared cultural and linguistic backgrounds. While the instruction of reputable experts is irreplaceable, insights provided by casual conversation with peers appear to have a longer shelf-life than dialogue consisting of terminology not yet mastered by students [1-4]. This raises the question: can rhetoric serve as an effective tool in student comprehension and retention of physics concepts?

If as a society our goals are to increase the workforce of the scientific community, improve scientific literacy, and diversify the scientific community's demographical makeup, we have work ahead of us [5]. While the responsibility of understanding science rests with the individual, the responsibility of clearly articulating scientific knowledge rests upon the shared scientific community. The gap between technical accuracy and public understanding could simply be a language barrier. Bridging this gap will move us closer to these worthwhile goals.

INTRODUCTION

The applied linguist Neil Murray defines technical language as "specialized terminology associated with a particular field or area of activity" [6] . The field of physics is replete with technical language. While useful for communicating complex concepts efficiently, this professional discourse can prove a barrier to students entering the conversation. The traditional approach to learning physics concepts is akin to an immersive language experience. Armed with a few compact lines of math and perhaps a sentence or two to define a term, students are promptly thrown into problems that require a variety of advanced skills. This can clearly be seen in problem sets, preparatory media for class participation, and even classroom instruction which all heavily rely upon mastery of technical terminology.

While this immersive approach does have the benefit of efficiently communicating ideas, it suffers from isolating those who have not yet mastered technical jargon. Problem solving, computation, and recognition of principles are combined with a need for understanding technical language. In opposition to this conventional structure, *Phsysics: A Treatise*—and its accompanying chatbot, Newt—have the aim of reversing this approach. This paper will explore the impacts of supplemental textbook and AI resources with specific stylistic design on self-selected students in an introductory, university level physics class through qualitative analysis of survey responses and grade distributions, as well as sensemaking of student engagement with these resources.

PROJECT OBJECTIVES

- Develop a supplemental text—with accompanying chatbot—as a resource for students in introductory physics classes targeting a personable authorial voice and reduced syntactical complexity.
- Pilot the text and chatbot in an introductory physics course.
- Examine the relevance of textbooks in an era of unprecedented AI progress.

RESEARCH QUESTIONS

Q1: What effect do the chatbot and textbook have upon post-secondary student achievement in the context of an introductory physics course?

Q2: What effect do the chatbot and textbook have upon post-secondary student perceptions of physics as a science in the context of an introductory physics course?

Q3: What do student preferences regarding the chatbot and textbook reveal about post-secondary student engagement, learning, and motivations?

LITERATURE REVIEW

The sentiment that the field of physics education needs reform is not a novel one [7]. Otero notes that "…despite more than a century of broad agreement among physicists on the type of instruction that should take place in physics classes, such instruction has yet to be achieved in the majority of US K–12 classrooms" (p. 54). Everyone, it would seem, agrees that teaching methods need to change to better accommodate and inspire students. Implementing such practices, however, has proven to be quite difficult.

The inception of this study was aimed at examining rhetoric and its effect on student comprehension and engagement. However, since the commencement of the project, a broad swath of the existing literature has made it clear that eliminating technical jargon from the delivery of physics concepts *does* have a positive, measurable impact on student performance [1-4]. The question has shifted from *if* rhetoric matters to *how* it can effectively be harnessed in the classroom.

In light of this evidence, this literature review serves two purposes. The first is to provide justification for the framework used in constructing the textbook using the findings of previous studies. The second is to explore the intersection between artificial intelligence as a classroom resource and student achievement.

JUSTIFICATION FOR FRAMEWORK

The organizational structure of hypotheses, assumptions, data, limitations and conclusions applies equally well to both scientific models as well as the texts used to explain them. Dr. Paul Strube, a respected lecturer, textbook author, and curriculum designer addressed this specific rhetorical design in his 1989 study "The Notion of Style in Physics Textbooks [8]. While much has transpired in the 40 years since the publication of this paper, Dr. Strube's criticism is as timely as ever:

> Since textbook writers appeal to a model of science that emphasizes logical, inductive, and deductive formal reasoning, the textbooks argue in the same way. The presentation to the reader follows… the classical model of the scientific method. There is no longer, however, a consensus view that this is in fact the way science does operate… Nevertheless, even if it is the way science does work, *it is not necessarily the best way for science education to proceed*. (p. 296) (emphasis added)

Hard sciences like physics having a clear modus operandi is insufficient justification for the delivery of that information to subscribe to the same pattern. The very fact that a degree in physics cannot be exchanged for a degree in communications should suggest that there are best practices exclusive to each discipline equally as well as the fact that a degree in physics cannot be exchanged for a degree in chemistry. Books are not atoms, and hence should not be subject to the methods used to develop an atomic model.

Progress towards any goal—pedagogical or otherwise—requires sacrificing old practices. The question then remains: why has the textbook genre not accepted these improvements? A number of theories could be posited, but the most plausible is that scientific convention has more metaphorical inertia than the Sun. Tradition is an entrenched part of any institution, and academia certainly falls prey to this. Beyond this, implementing course corrections is extremely costly: vast amounts of time, money, and training are required to deploy different practices effectively *after* they have been found in the literature (a lengthy process in and of itself). All of this comes without even beginning to consider the hefty bureaucratic barriers within educational institutions. In short: change is very hard.

Furthermore, the growing body of research on interleaved practice suggests that a pedagogy which structures concepts into compartmentalized, modular forms—while more traditional and comfortable—is not as effective as study structures that span multiple topics and types of problems, so long as the topics are within the same field [9-15]. These findings suggest that students should be synthesizing the information they learn in class with their current understanding of the broader world.[i] Samani, for instance, found in his 2021 paper published in *Nature* that interleaving concepts moved their experimental group's initial median test scores up by 50% compared to the control group. This is not a small effect. One, it is posited, that could reasonably be applied to the *delivery* of course content alongside the content itself. Of course, care must be taken to not view interleaving as a panacea, for Richter is quick to note that despite ample evidence for the benefits of interleaving, there is significantly less consensus upon how to implement this strategy. While this admission appears to temper the vitality of these claims, the very same study goes on to note in its concluding remarks that further experimental research designed around the implementation of interleaving will be integral to finding a solution. This and other open ended calls to further research is ample provision to conduct a study such as is presented in this very paper.

A fair criticism of this rhetorical design is the notion that if technical vocabulary is *not* introduced and used extensively, students will never develop the language of professionals. This criticism has merit. However, the framework of these resources is not to completely exclude scientific terminology: it is simply to give concepts equal footing with their descriptions. Aside from this, eliminating technical vocabulary would essentially be a re-mixed version of modular teaching, invalidating the entire argument to synthesize learning. An extensive study in elementary education concluded that introducing precise mathematical language *before* asking students to do math would improve student achievement [16]. Familiarity with a toolbox allows a skilled laborer to immediately recognize the right wrench for a stubborn bolt. Familiarity with concepts allows a theorist to communicate nuance and detail clearly.

21st century problems require complex and multifaceted solutions. As the contemporary buzzword, climate change is an exemplary problem that demonstrates this need. No one discipline, nation, or technology can counter this pervasive, global issue single handedly. As the English poet John Donne penned: “No man is an island entire of itself,” [17] and nowhere might this prove truer than in our shared responsibility to process scientific data into digestible models and calls-to-action for an audience broader than our individual fields of study. If the timescale of discourse surrounding physics education reform has taught us anything, it is that identifying solutions is not enough: we must find ways to appropriately enact solutions. This will require that we know concepts as well as—if not better than—the vocabulary used to describe them.

Surprisingly, engagement with course material can be greatly impacted by providing engaging material. Dr. Jan Packer of the University of Queensland expertly notes that learning is by definition an exploration of the world that allows participants to willingly open themselves up to new ideas and perspectives [18]. Discovery is integral to learning and is applicable to a broad degree. “The design of exhibits [like Disneyland] in this style of theme park is such that, incidentally, the visitor learns a great deal of information about a range of issues…” [18]. It could easily be argued that no one goes to Disneyland to learn, and yet they end up doing so anyways. When engagement is a primary design element, education is a natural byproduct. Thankfully, education has a significant amount of potential to be engaging [19]. Cardinot’s study followed the usage of a board game used to teach an astronomy class. The discussion suggests that enabling students to be active participants in the learning process is critical to student motivation and performance. Of course, such a design needs to be optimized to fully reap the benefits.

The idea that education can simultaneously be effective and entertaining is shared by academics directly within the field of physics as well. Surveys and interviews among 26 top Czechoslovakian researchers in 2023 verified that interesting course material should be a top priority for physics curricula [20]. Discussion in this study found that many professionals heavily emphasized aspects of teaching that are traditionally associated with humanistic curricula such as student interest in course material, stories, and the open-ended nature of physics as a science. While the study could not definitively speak to the motivation for these remarks, some interviewees critiqued traditional teaching methods, associating the model with student dissatisfaction with physics. In other words, these researchers emphasized the significance of discovery and engagement in physics classes. When stopping to consider that these professionals consistently encounter discovery, engagement, and problems without clear solutions in their day-to-day work, this connection seems quite clear. Žák further emphasis in his review of the literature that "the selection and ordering of topics within the physics curriculum [is] being guided more by habit or tradition than by cognitive research or sound pedagogy." This corroborates the argument made earlier that there is a wide gap between educational understanding and educational practice.

Table 1. Definitions of terms in Dr. Strube's analysis of textbook style.

Term	Definition	Traditional Example	Counter Example
Authorial Voice	The tone and style in which a text is written.	"Experimental measurements of our Sun suggest that we receive an average of 1.36e3 Watts/m^2."	"Earth gets roughly as much sunlight as taking each bright spot in *Starry Night* and replacing it with a 75 Watt light bulb for every area the size of the painting."
Precision	Providing rigorous definitions in unambiguous terms.	"A vector is a rank one tensor with invariant magnitude under coordinate transformation."	"A vector is an object in math that can be used to describe orientation and strength."
Reduced Context	Constraining a concept or principle to the discussion at hand, without consideration of other applications or interpretations.	"The *k* vector describes the wavenumber in each mutually orthogonal direction."	"The *k* vector appears in optics, quantum mechanics, and mathematical physics."
Limited Syntax	The structure, length, and complexity of sentences.	"The non-linear torsional mode creates a sheer stress within the crystal which under conditions of resonant oscillation leads to a divergent magnitude of force, thus resulting in the splitting of the lattice."	"As torque is applied, the crystal is strained. A sheer force is generated. The effects are non-linear. At the proper frequency, the force diverges. Eventually the lattice splits."
Rhetorical Model	The flow of an argument, which is generally "linear" in the sense that arguments progressively build upon one another.	"Solar sails rely upon radiation pressure. Photons carry momentum. Momentum imparted on a surface produces a force. This force generates an acceleration, and the mass moves."	"Photons carry momentum. Solar sails rely upon radiation pressure. This force generates an acceleration, and the mass moves. Momentum imparted on a surface produces a force."

Clearly, materials with an accessible voice are well supported by the existing literature. To move forward, we must have clear definitions to work with. Dr. Strube (1989) provided a thorough classification of academic formalism which we adopt in this work. His exhaustive classification of formal textbook tone includes the following five terms, described in detail within Table 1. To demonstrate the range of possibilities available for each of these rhetorical devices, both a traditional and non-traditional "counter example" are provided for comparison.

Noting that "[t]here is no discipline-bound need for science textbook writing to be either impersonal or syntactically complex," a primary goal of this experiment will be to intentionally challenge the traditional perspective on authorial voice and limited syntax as described in Table 1. This coincides directly with the motivation, project objectives, and the existing body of literature, as has been examined in this review.

CONTEMPORARY RESEARCH

Being an emerging field of research, there is still much to be learned regarding the intersection of generative AI and education, but the concept of introducing "smart" technologies into the classroom setting has been around for decades [21-24]. These various studies have examined the impact of virtual reality to mass-produced smart assistants, such as Amazon's Alexa. An uncharted intersection does exist, however, in asking how the structure of AI responses (including pedagogical structure and style) impact student's academic performance and perception of class material. This gap exists particularly for postsecondary education, which has been explicitly called out in the literature as a point for future studies [25].

With the current state of generative AI, response length and detail are a major concern. Armed with a tool that has the potential to describe nuanced processes and supply numerical solutions to complex problems, the nature of student learning shifts significantly [26]. Generative AI is out of its natural context in providing educational instruction, since the major goal of large language models is to provide information, not to employ educational tact in assisting the user's learning. Abdelghani specifically notes that major inroads will likely be created by training generative AI models on texts that already contain pedagogical formalism so that these results can be "baked-in" to the responses [26]. I address this specific concern in the development of the materials section below with the training resources used to build Newt.[ii]

METHODS

SCOPE & LIMITATIONS OF STUDY

This project and its analysis were conducted during the Fall 2024 semester of Brigham Young University. As such, it is confined to the particular demographic (see section below for more details) of the institution and further limited by the students who not only registered for a general course in physics, but also those who self-selected to participate in the study.

As students self-selected to participate in this research study, there is inevitable potential for self-selection bias. To counteract this bias, the entire class was given midterm 1 prior to receiving any of the experimental materials. Although no bias can be fully eradicated from a given study, the design of this study using the first midterm as a benchmark does mitigate the effect and provide a clearer lens through which to view the results.

In addition to these items, tracking of material usage was limited, and thus there may have been students who actively used the resources *without* participating in the study, though given the level of engagement recorded from the chatbot, this is not very probable.

DEVELOPMENT OF MATERIALS

Although many scientific authors have written works that incorporate light-hearted elements in informative works (see for example Randall Munroe's *What If?*) [28], these texts are less likely to serve as material for classroom instruction. A major lodestar for the stylistic design of the course resources was James Kakalios' *The Physics of Superheroes* [29]. The primary design element of these resources is accessibility for an audience not directly inclined towards the natural sciences. Although a major motivation for implementing the textbook and chatbot was to test the effects of non-conventional authorial voice and reduced syntactical complexity, this was not a feasible aspect to test for given the constraints of the study. Despite these limitations, the rhetoric of the textbook and chatbot were centered on this guiding philosophy.

Phsysics was intentionally designed with a non-technical audience in mind. The language used in any written work certainly affects it's demographical reach. Thus, to create an inviting format that is appealing to a larger audience, *Phsysics* was written with a "concept-first" approach. This approach does not preclude the use of more formal definitions, but it does avoid discussing concepts in a manner which utilizes these formal definitions. See Appendix A for a few examples of this style in action.

The chatbot was fed *Phsysics* as well as the first 6 chapters of OpenStax's textbook on university physics with specific instructions to generate text following the tone and style of the textbook developed for the course. In addition, several engineered Q&A sections were provided for the chatbot as summarized in Table 2 on the following page.

Table 2. Q&A responses given to chatbot to assist in training.

CHATBOT PROMPT ENGINEERING		
Question Posed to AI	**Response**	**Purpose of Instruction**
What should the length of my responses be?	“Limit your responses to be 250 words or less in all cases.	Avoiding student passivity, and “brevity is the soul of wit.” [30]
What is my goal as a chatbot?	“Answer student questions clearly and cleanly. Do not provide an answer if you do not know how to respond.”	Constrain Newt to only respond with confident answers.
How should I, as a Chatbot, sound?	“Use the style, tone, and voice of the attached text, Phsysics as much as possible. Be rigorous in providing accurate descriptions. Be approachable and conversational.”	Constrain Newt to operate within the same stylistic framework as the textbook.
How much information should I provide a student asking questions?	“Your purpose is to prepare a student to move forward to the next step in their understanding. Your purpose is not to do all of the work for a student.”	Constrain Newt to provide support and encouragement, rather than simply giving students answers.
What kinds of responses should I give to students?	“Whenever possible, respond to a student’s inquiry with a question, a relevant principle to solving the problem, or setup for the next step in solving the problem.”	See above.
Can I simply give students the final answer to a problem?	“Without exception, no.”	See above.

COURSE DESCRIPTION

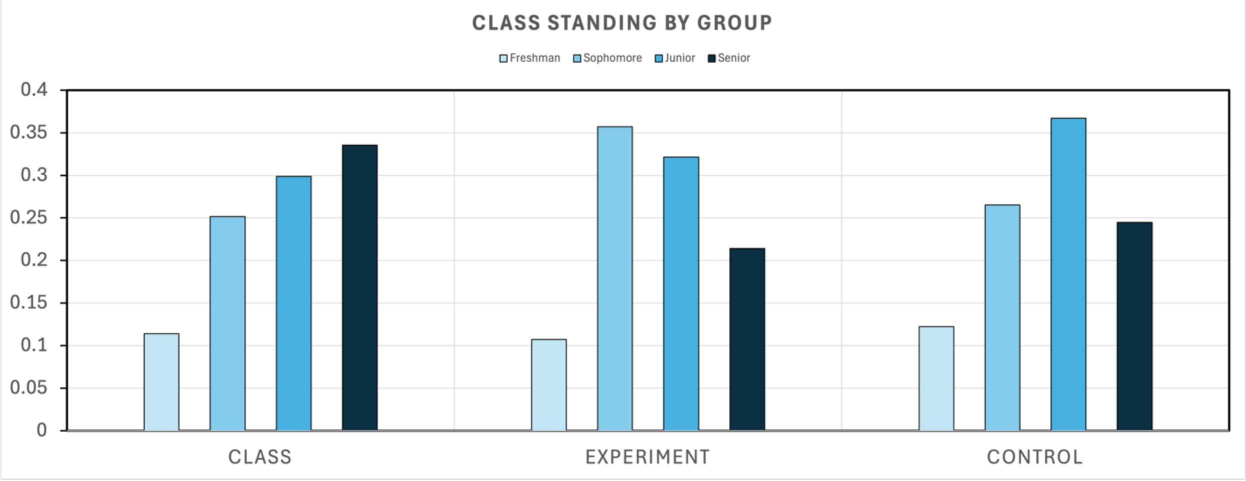
PHSCS 105 is a lab and lecture based applied physics course which does not require calculus. Mechanics, heat, wave motion, and sound are the primary subjects of interest for the course. Heavy emphasis is placed upon solid comprehension of course material, which is measured primarily via examination and homework sets (for a combined 90% of the class grade).

The textbook and chatbot were introduced to Dr. Adam Bennion's Physics 105 class during the 2024 Fall semester at Brigham Young University. While this choice does restrict the sample size, it also ensures that the individual teaching the course is less likely to serve as a confounding variable. The section of Physics 105 being followed in this study had a population size of $N_p = 298$. The self-selected sample from this population was $N_s = 28$.

COURSE DEMOGRAPHICS

The data reported in this section of the report serves two purposes. The first is to compare the population of the class to the self-selected sample group and the control groups. The second is to ensure full disclosure of the study for future research to understand the context in which these findings are based. Details for each group are provided in the subsections below.[1]

Figure 1. A breakdown of the class standing for each group under consideration in the study.



[1] It may prove beneficial for the reader to note that color-coordination has been kept consistent between section graphs, so that each major and year in school pertains to one color that is maintained throughout the following subsections. Furthermore, the group labeled "Other" contains majors with ≤ 10 students representing the major while the group labeled "Individual" is reserved for majors which only had 1 student with this major in the class.

Figure 1 is a relative breakdown of the class standing between the population, the experimental sample, and the control groups. Due to the high volume of majors in the class, an equivalent general representation for majors is visually unwieldy and does not reveal any particularly interesting trends.

Population Breakdown

The class consisted of $N_p = 298$ students, with 43 majors of the 198 undergraduate majors offered at BYU (a total representation of ~22% of the university's programs). Among these 43 majors, only 7 were represented by 10 or more students (or ~16% of the majors present in the class). Furthermore, not a single student in the entire population was in a major even relatively adjacent to physics. This is not surprising given the fact that this particular course track consists of generals designed to meet physics requirements for non-physics majors.

Experimental Group Breakdown

The experimental group is—of course—of special interest, and there are several unique features pertinent to this group which will be examined in detail.

Figure 2. Breakdown of majors and class standing for PHSCS 105.

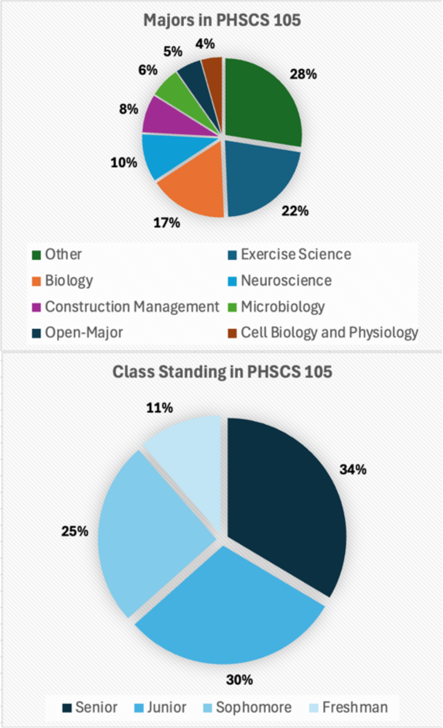
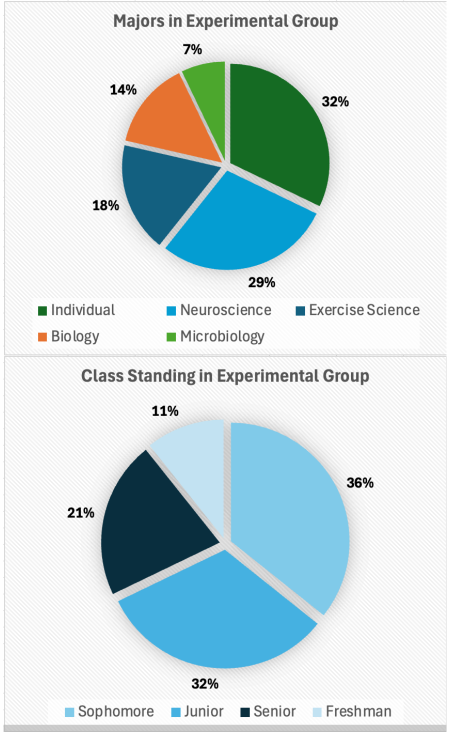
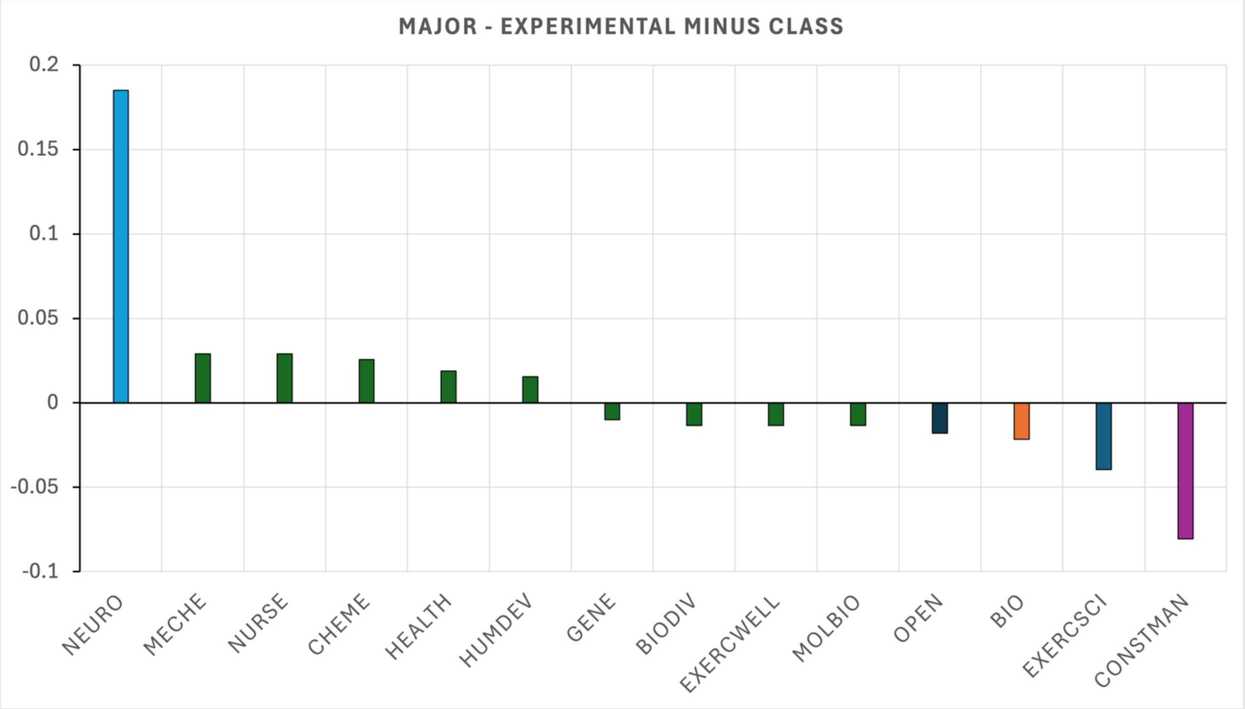


Figure 3. Breakdown of majors and class standing for the self-selected experimental group.



It is particularly interesting to note that the top four majors in the population (accounting for 77% of the class) are the same four majors present—albeit in a different order—in the experimental group (accounting for 93% of the experimental group). This is more clearly visible in Figure 4.[2] The trend shows that a disproportionate number of neuroscience majors self-selected to participate in the study while a disproportionate number of construction management majors chose not to participate. This would seem to suggest that the aim of reaching students who are not in fields closely related to STEM was not accomplished by allowing for self-selection.

Figure 4. A summary of the difference in percentages for the key majors in the experimental group compared to the class population. Each percent in the experimental group was subtracted from the corresponding major in the population.

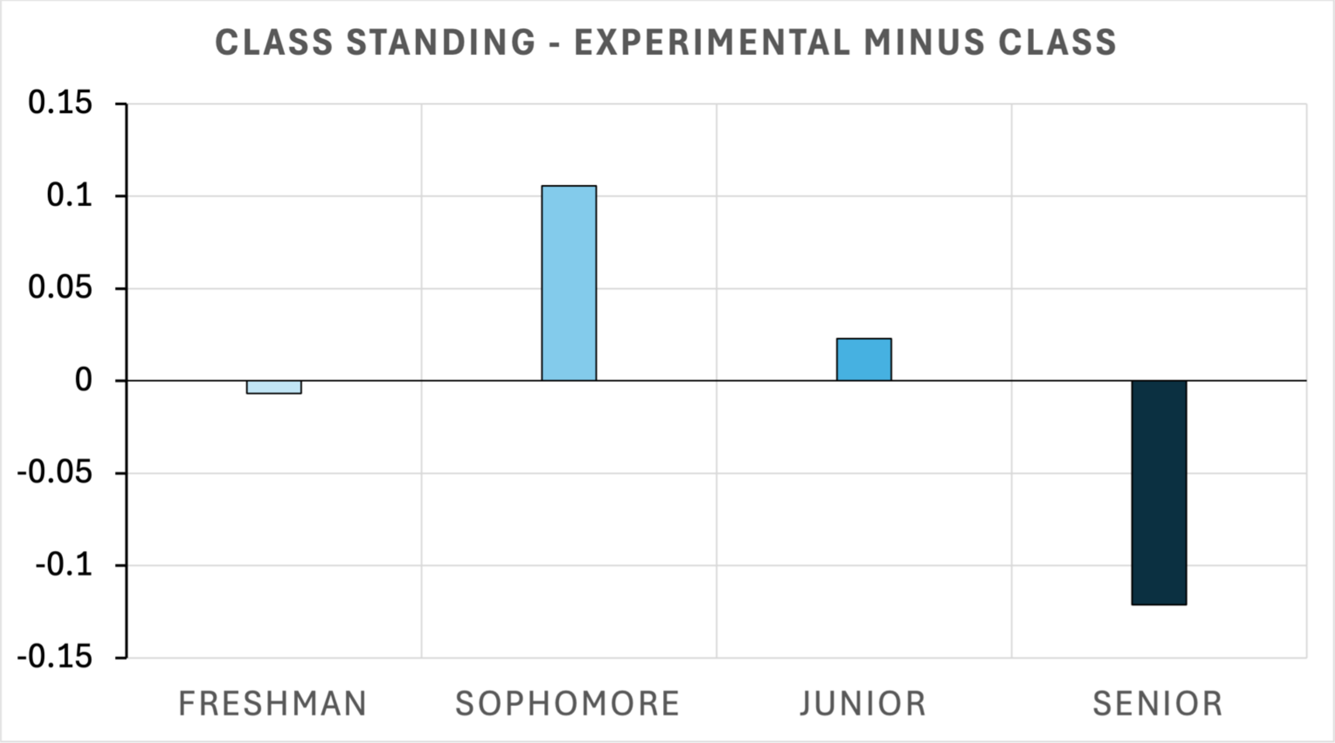


Another phenomenon occurred in regard to class standing. Although the percentage of Freshmen who participated was virtually identical to the class population, significantly more Sophomores and Juniors participated while a disproportionate number of Seniors opted out of the

[2] It is important to note with this figure that only deviations of ≥ 1% have been plotted. Also, majors have been abbreviated.

study. This bias likely also comes from internal motivations as well, though significantly different in nature from the STEM motivation considered earlier. These differences are summarized in Figure 5.

Figure 5. A comparison of class standing between the experimental group and the class population. Each percentage in the experimental group was subtracted from its corresponding percentage in the population.



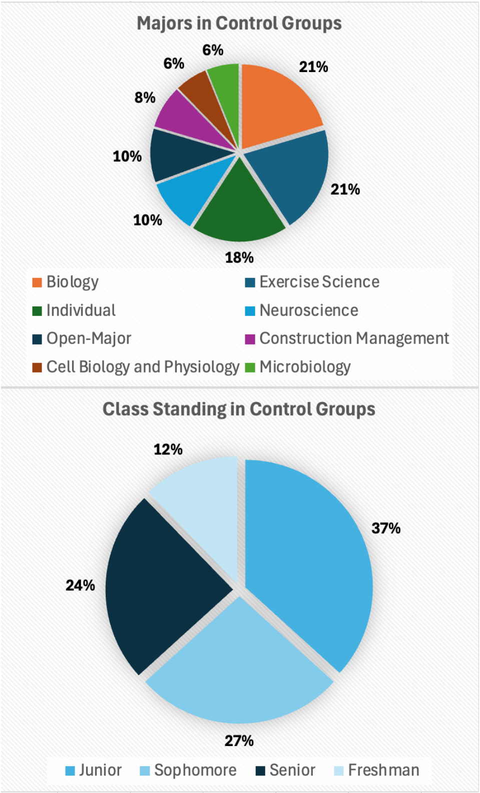
Control Group Breakdown

Due to random sampling with replacement—including for members of the experimental group—there were only 49 students surveyed total between the two control groups. The breakdown in Figure 6 combines both control groups and only counts individual students once.

DATA COLLECTION

Collecting a wide variety of data ensures a more complete characterization of a population. This being the case, aggregate data on demographics, majors, and credit hours are reported in the section "COURSE DEMOGRAPHICS" above. Scores for midterms 1 through 3 were collected from the class population randomly with replacement to provide two control groups of equal size to N_s for comparison. Furthermore, scores for midterms 1 through 3 and survey data were collected from the experimental group.

Figure 6. Breakdown of the majors and class standing of the students in the randomized control groups.



The survey data was collected using three different Google forms structured on the 5-point Likert scale. The questions regarding student perceptions and engagement can be found which can be found within the "RESULTS AND DISCUSSION" section below. This data was further analyzed using standard statistical methodology to present the results found in this paper.

In addition to these data points, student interactions with Newt were also collected and analyzed for sensemaking. The collected interactions run from October 5th, 2024 until November 5th, 2024 (32 days total).

DATA ANALYSIS

In order to perform a two-tailed T test, several conditions must be met. The two-tailed T tests were applied only to the midterm scores, and thus only this data must meet these conditions. These conditions are justified as follows:

Continuous Data Set: The scores of a midterm in the many combinations of point values (particularly for free response materials) so as to be effectively a continuous variable. Counting half points and integer scores means that the data is a set with at least 201 discrete points.

Random Sampling: While the sample was self-selected, the control groups were selected randomly with replacement. Furthermore, analysis of class demographics demonstrates that a representative sample of the population was selected for the control groups.

Homogeneity of Variance: 28 cases and 3 groups necessitate comparison with $F_{2,25} = 2.52831$. Calculating the Levene test statistic for these groups gives a value of $W = 0.76$ for Midterm 1, which is clearly beneath the critical value. Thus, we accept the null hypothesis that these groups have homogenous variance.

Normality: The Shapiro-Wilk test for the control groups is an order of magnitude beneath the desired 0.05 result for Midterm 1. Despite this finding, (which is clearly a result of the hard 100% cap on tests) we proceed with analysis in good faith since treating non-Gaussian distributions like they are normal is common practice in research.

Independence: We assume that the students are free agents whose scores cannot be predicted based upon the scores of their classmates.

Sample Size: Using an average standard deviation of the three midterms for the experimental groups yields a standard error of 3.6. While a larger sample is always preferable, the group sizes are large enough to notice any statistically significant differences.

Each midterm was then analyzed to obtain its t test statistic:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

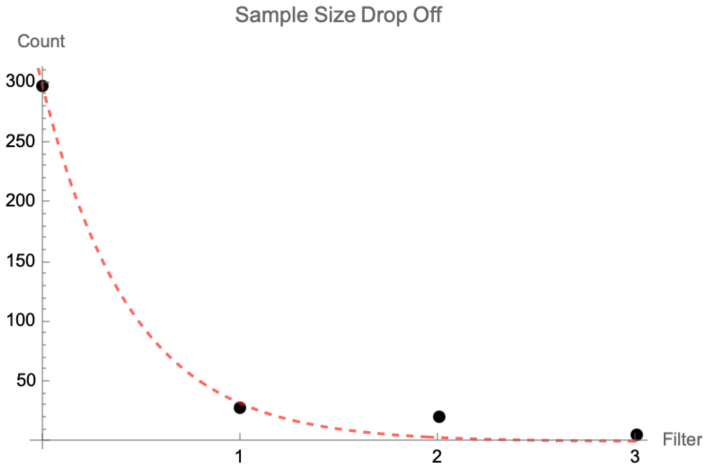
Further analysis was conducted on the data sets presented in this paper with specific details presented in each individual case where further extrapolation is required to fully appreciate the results. Beyond this, standard statistical methods such as those found in an introductory statistics text like normalization, data visualization, and modeling were the only tools used for sensemaking in this report.

RESULTS AND DISCUSSION

SAMPLE SIZE

Throughout the course of the study, the sample size of respondents decreased dramatically. While this is an unfortunate event for the analysis of data in this particular study, it is in and of itself a useful trend to characterize and understand. Figure 7 highlights this trend. Each "filter" is an activity that the volunteers were asked to participate in. Filter 1, which produced $N_s = 28$ was the signing of the research consent form. Filter 2 was the completion of the first general survey sent to the experimental group, giving $N_s = 21$. Filter 3 were the follow-up surveys at the conclusion of the study, giving $N_s = 6$.

Figure 7. A scatterplot (with exponential model) demonstrating how repeated requests for engagement with the study impacted the study's sample size.



PERCEPTION SURVEY RESPONSES

The class responses to the surveys examining student perceptions of physics are summarized in the tables below. Each question has been assigned a color. The intensity of the color corresponds to the percentage of a given response.

Table 3. Summary of the distribution of responses for the initial survey.

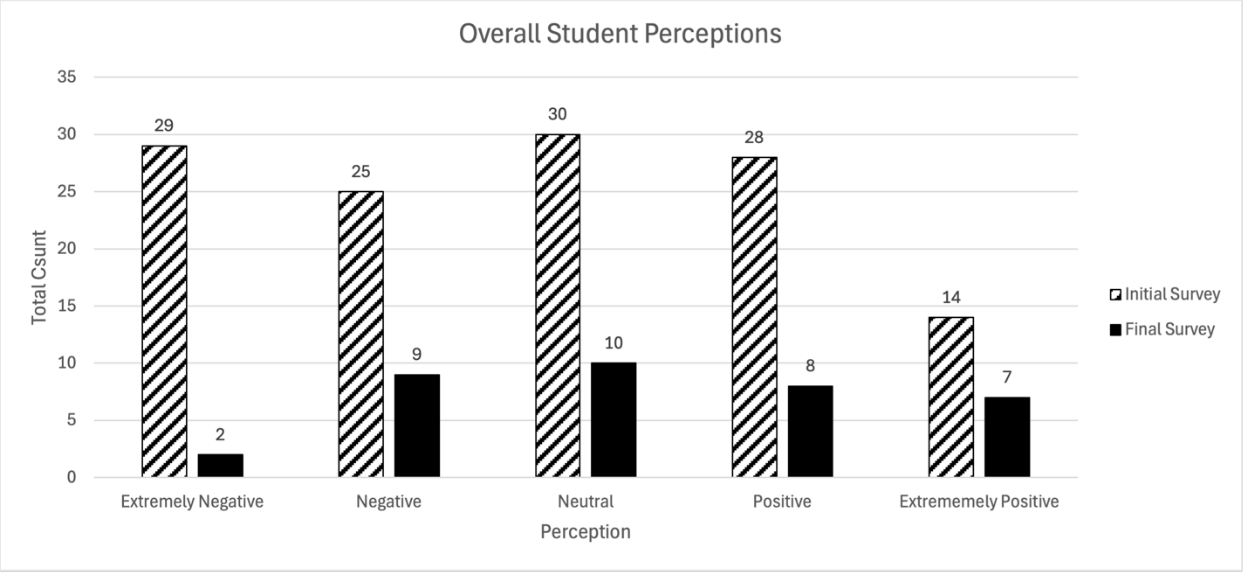
Question Connotation	Positive	Negative	Negative	Negative	Negative	Positive
Initial Survey N = 21	"**Anyone** can learn physics."	"Physics is inherently difficult to learn."	"It is impossible to present physics concepts in an intuitive manner."	"I struggle understanding what certain words or phrases mean in physics."	"Physics is boring."	"I would consider pursuing a degree in physics."
Strongly Agree	33.3%	28.6%	4.8%	19%	14.3%	0%
Agree	38.1%	42.9%	23.8%	19%	14.3%	4.8%
Neutral	28.6%	28.6%	33.3%	28.6%	19%	4.8%
Disagree	0%	0%	33.3%	23.8%	33.3%	19%
Strongly Disagree	0%	0%	4.8%	9.5%	19%	71.4%

Table 4. Summary of the distribution of responses for the final survey.

Question Connotation	Positive	Negative	Negative	Negative	Negative	Positive
Final Survey N = 6	"**Anyone** can learn physics."	"Physics is inherently difficult to learn."	"It is impossible to present physics concepts in an intuitive manner."	"I struggle understanding what certain words or phrases mean in physics."	"Physics is boring."	"I would consider pursuing a degree in physics."
Strongly Agree	66.6%	0%	0%	0%	0%	0%
Agree	16.7%	66.6%	0%	16.7%	0%	0%
Neutral	16.7%	33.3%	50%	50%	16.7%	0%
Disagree	0%	0%	50%	16.7%	50%	66.6%
Strongly Disagree	0%	0%	0%	16.7%	33.3%	33.3%

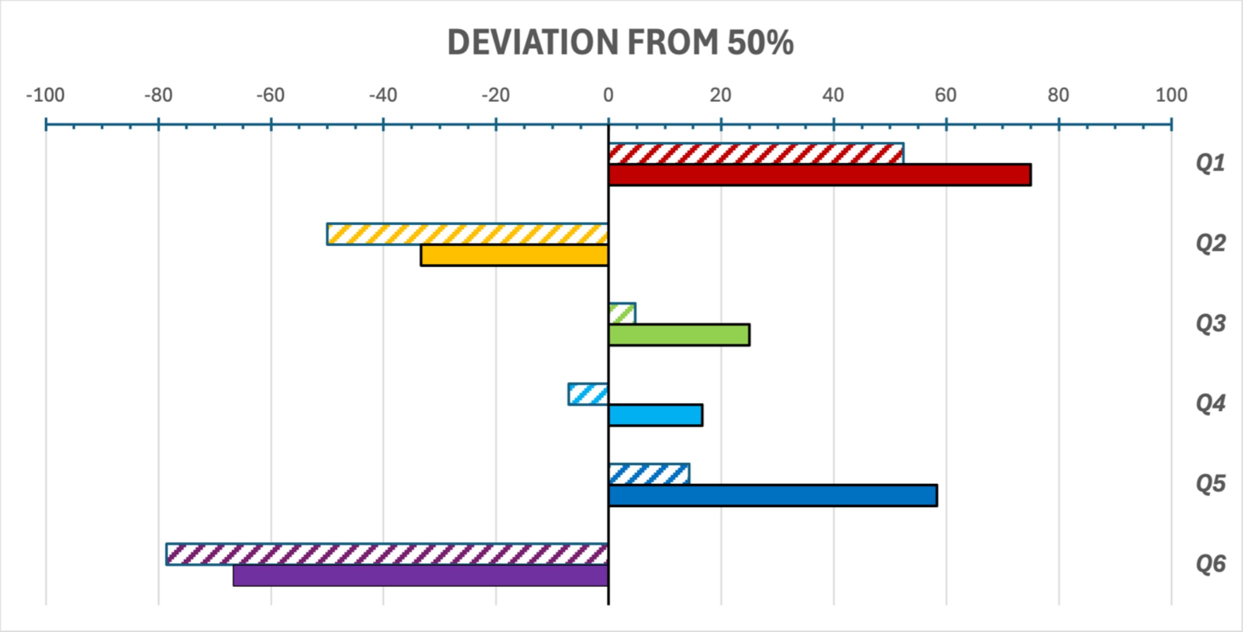
By using the Likert scale, the results of all six questions were combined to form a bar chart demonstrating the overall perception of the students. For example, all of the individual "Strongly Agree" responses were added together and counted as a single value. However, since each question has a built in perception, those questions where a "Strongly Agree" meant that the student had a *negative* perception of physics has their values swapped so that the overall distribution has positive values to the right, negative values to the left, and zero at the center. This coarse metric obviously has limitations, particularly with its resolution of detail, but it is nevertheless a reasonable gauge of the participants overall perception of the science as a whole.

Figure 8. A tally of the overall Likert scores for the initial survey and the final survey.



In addition to this metric, a measurement of conviction can be made by comparing the sum of the received scores on the Likert scale to a fully positive response (100% of respondents "Strongly Agree") to a fully negative response (100% of respondents "Strongly Disagree"). This was achieved by measuring each scores deviation from neutrality (50%) of the possible total score. Once again, questions with a negative connotation have been reversed so that a negative value truly aligns with a negative perception. It should be noted that the initial survey sample consisted of $N_s = 21$ student responses, while the final survey consisted of $N_s = 6$ student responses.

Figure 9. A measure of the participants deviation from neutrality in both surveys. In this figure, the striped bars represent the initial survey while the solid bars represent the final survey.



The change in percentage from the initial survey to the final survey is summarized in the table below.

Q1	Q2	Q3	Q4	Q5	Q6	Average
+ 22.6%	+ 16.7%	+ 20.2%	+ 23.8%	+ 44.0%	+ 11.9%	+ 23.2%

Table 5. A summary of the change in deviation from neutrality. All of the scores were positive, indicating that the distribution shifted towards positive views for each question.

To further explain, each survey has a minimum value of N for each question and a maximum value of 5N. The mean value of these is 3N with 2N values above and below this value. Thus, we can measure a given deviation from 50% by subtracting the sum of the scores by 3N and dividing by 2N on either side, giving us a range of values from -100% to 100%.

MATERIAL USE SURVEY

The material use survey presented some of the most fascinating trends. In total, five different questions regarding each of the supplemental materials revealed starkly different opinions regarding the textbook and chatbot. The substance of these five questions is as follows:

1. How often did you consult this supplemental resource?
2. Was the supplemental resource useful for learning concepts?
3. Was the supplemental resource useful for completing homework?
4. What aspects of the supplemental resource did you find appealing?
5. How would you categorize the tone and style of the supplemental resource?

We examine each in turn.

Q1: Resource Consultation

Strikingly, those who chose not to engage with the resources were the same across the board: the percentages for students who did not use either resource (42.9%) or only used the resources once (14.3%) was the exact same for both textbook users and chatbot users. Among those who chose to use the resources, the difference was simply in degree.

While chatbot users were more nuanced in the frequency with which they used Newt, the broad strokes of data indicate that on average textbook users only opened their resource 1-2 times per week. In contrast to this, those who engaged with Newt did so with at least twice the frequency of textbook use, with many reporting 5 or more uses per week.

The fact that Newt was consulted more often than the textbook is not a surprising result; indeed, to some degree it was anticipated. Unfortunately, the metric caps at 5+ uses per week, so finer detail is not available. What *is* striking is the result that students *reported* being twice as likely to consult a chatbot over a traditional textbook. Refined data on the likelihoods of choosing between resources of this nature would be an excellent candidate for a metric in future studies. The poll data available here does not provide a statistically significant figure to make claim as to this likelihood, but it is suggestive of this possibility.

Q2: Resource Utility (Learning)

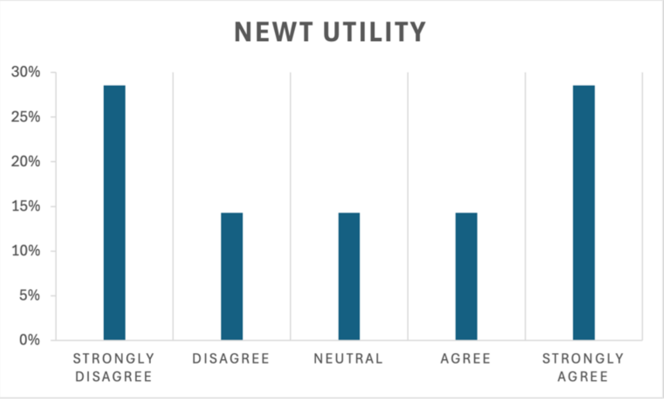
Not a single student reported having a positive perception of the textbook in this regard: the distribution solidly fell at or beneath neutrality. Conversely, student's self-reported experience with the chatbot revealed a unique and divisive distribution. This distribution is discussed further in the next section as it is literally indistinguishable from the distribution for the chatbot in Q3.

Q3: Resource Utility (Assignments)

Peculiarly, the distribution for the textbook swayed slightly here, indicating that while no student felt that the textbook taught anything well (see Q2), two students agreed that it assisted in completing assignments. The author infers that this may simply be a nicety of the students producing such an anomaly, although with the sample size it is difficult to draw many conclusions.

As mentioned above, the distribution of responses for Newt’s utility in completing homework assignments was the *exact same* as the one found in Q2. This may not be entirely unusual considering how similar Q2 is to Q3, but the distribution itself is quite interesting. Figure 10 displays this double-distribution, which—if anything—is clearly not a normalized Gaussian curve. This division is striking, and the repetition is suggestive that opinions over the utility of AI as a classroom tool vary on a continuum.

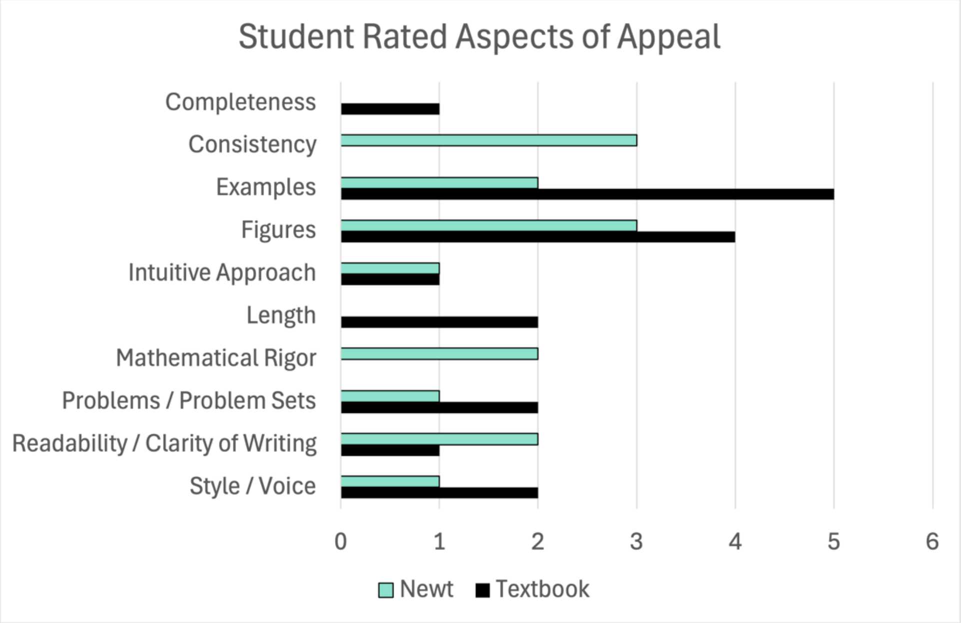
Figure 10. The distribution for Q2 and Q3 regarding Newt's utility.



Q4: Preferred Aspects in Resources

Figure 11 summarizes the student’s reported aspects of appeal for both the textbook and Newt. It is curious to note that in both cases, style and readability were not highly ranked.

Figure 11. A summary of the resources most appealing elements, as reported by the student users.



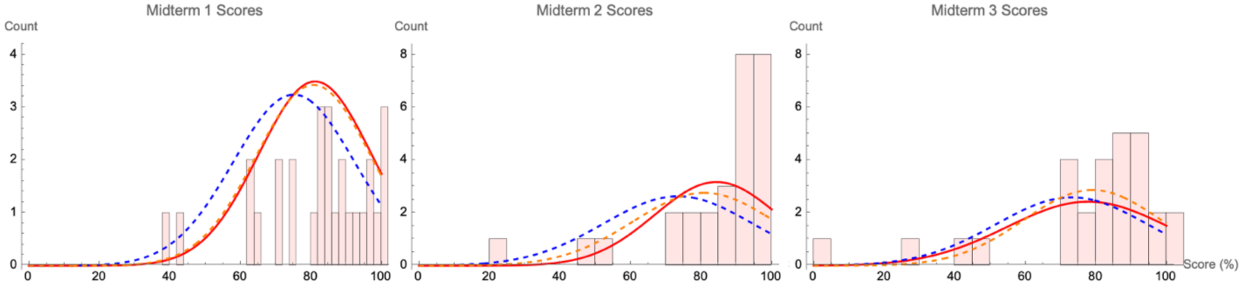
Q5: Resource Tone and Style

The data from this question indicates that the students recognize the intended style of the resources. The majority classified both resources as “Casual / Conversational.” This indicates that whether or not students like the pedagogical design of the resources, it was certainly sefl-evident.

GRADE DISTRIBUTIONS

Test scores in the class are summarized in the following normalized distributions (note that the distributions have been multiplied by the total integrated area under the histogram curve to scale visually). For statistical analysis and visual representation, the Gaussian distributions of two $N_s = 28$ control groups randomly selected from the population are also displayed. For each graph, the solid red line represents the experimental group. A frequency histogram with a step size of 5% is also shown for the experimental group. The blue and orange dashed lines represent the two separate control groups for each test.

Figure 12. A summary of all three midterms for the two control groups and the experimental group.



The statistical values pertaining to these distributions are summarized in the table below. At an α value of 0.05, the critical t-value is 2.052 with 27 degrees of freedom.

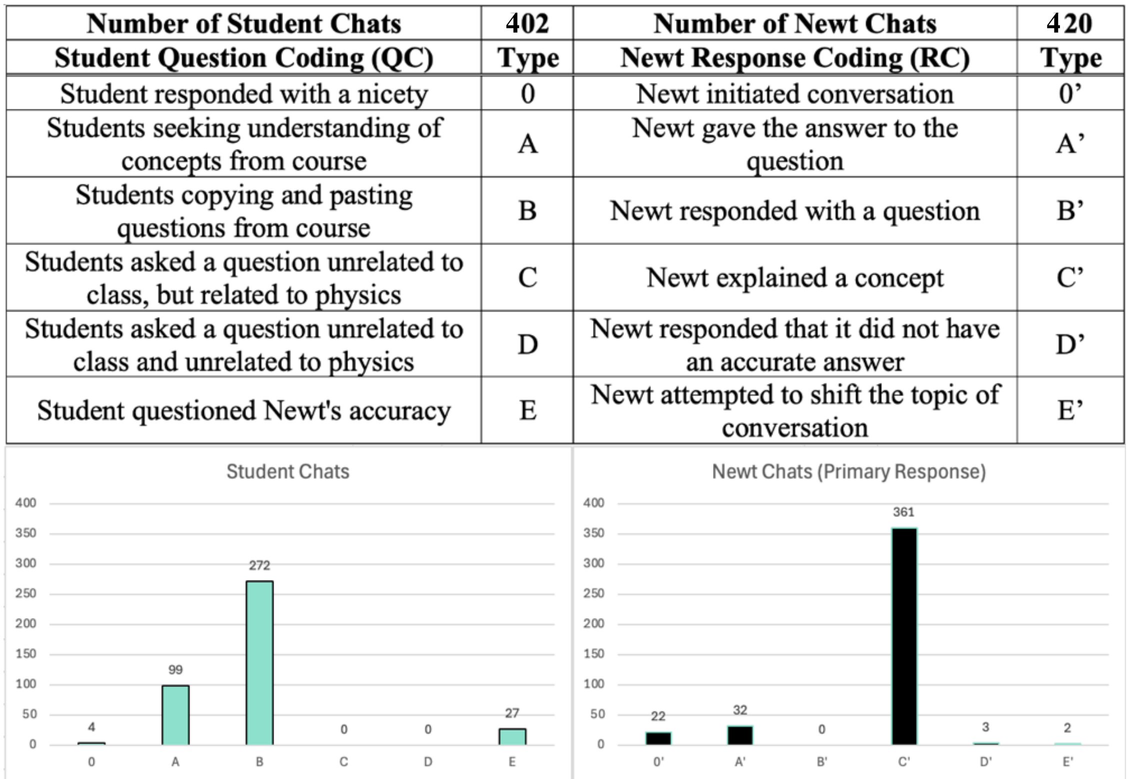
Table 6. A summary of the statistical values associated with the distribution of test scores for midterms 1 through 3.

N = 28	Type	Mean	Standard Deviation	Standard Error	t-value	p-value	Significant?
Midterm 1	Experiment	81.16	15.98	3.02	N/A		
	Control 1	75.14	17.20	3.25	1.36	0.19	No
	Control 2	80.71	16.28	3.08	0.10	0.99	No
Midterm 2	Experiment	84.36	17.57	3.32	N/A		
	Control 1	73.46	21.18	4.00	2.09	0.05	Yes
	Control 2	80.82	20.19	3.81	0.70	0.49	No
Midterm 3	Experiment	77.75	22.96	4.34	N/A		
	Control 1	73.14	21.50	4.06	0.77	0.45	No
	Control 2	79.0	19.43	3.67	-0.22	0.83	No

CHATBOT INTERACTIONS

Student interactions with the chatbot were varied and—on occasion—novel, but many of the interactions fell into the major buckets of students copying and pasting questions from the course (~68% of student chats) with Newt providing a detailed explanation of working step by step

Figure 13. A summary of the chats for both students and Newt, with the coding key applied to each interaction.



Number of Student Chats	**402**	**Number of Newt Chats**	**420**
Student Question Coding (QC)	**Type**	**Newt Response Coding (RC)**	**Type**
Student responded with a nicety	0	Newt initiated conversation	0'
Students seeking understanding of concepts from course	A	Newt gave the answer to the question	A'
Students copying and pasting questions from course	B	Newt responded with a question	B'
Students asked a question unrelated to class, but related to physics	C	Newt explained a concept	C'
Students asked a question unrelated to class and unrelated to physics	D	Newt responded that it did not have an accurate answer	D'
Student questioned Newt's accuracy	E	Newt attempted to shift the topic of conversation	E'

through the problem to provide a numerical answer with units (~86% of Newt chats). These findings are summarized in Figure 13.

The type of responses Newt is prone to give varies significantly depending upon the type of question posed. In instances when students asked legitimate questions seeking understanding of the course material (type A), Newt often gave concise, conceptual responses that clarified ideas, and/or discussed the significance of pertinent equations. In instances where students copied and pasted questions directly from the course website (type B), however, Newt took the reins and left little room for problem solving as it worked problems out step-by-step, often arriving at precise numerical solutions complete with units. When data necessary for solving the problem was missing, it generally had no impact upon Newt's responses, instead throwing out symbolic solutions ready to plug values in.

Levels of critical thinking and interpretation of results was not absent in the chats, but by no means the major interaction between the students and Newt. Even being generous and combining the students legitimate questions (type A) with their questioning of Newt's methods or results (type E) only accounts for 31% of messages sent by the students. Even at this, students often remarked to Newt that "the correct answer was [NUMERICAL VALUE HERE]." This suggests that students were not actively questioning the chatbot, but rather funneling the correct answers from automated submission feedback and attempting to have Newt rectify the error. It must be noted, though, that there were instances where students would question Newt's methods and ask why the answer was not something else; this most commonly occurred on multiple choice or conceptual problems.

The absence of critical thinking in interactions was by no means one sided though. Newt frequently made unfounded assumptions to complete a problem, and less frequently recognized or noted the assumptions as being potentially detrimental to the solutions. In all fairness to Newt, though, many students asked questions about visual images or simply did not provide pertinent data for a given problem that was necessary: it is hardly surprising that a generative AI would then make-up something, particularly when it has been instructed to generate content to assist students with problem solving. Curiously, Newt struggled with computation in a variety of ways. When students responded that Newt had given a faulty result (type E), the chatbot would run through the problem from the start and either arrive at the correct answer, get a different wrong answer, or simply change the value of one of the system's initial parameters—changing a given to a variable to solve for—to arrive at the correct answer. Despite these drawbacks, students repeatedly gave Newt numerical values to calculate an answer from a previously given symbolic solution rather than plugging these values into a calculator themselves.

Although Newt's effectiveness is up for question, students certainly engaged with the bot to a significant degree. Whether this was because the chats were useful, or students were willing to gamble their grade on its answers is currently unclear from the data. In any case, there certainly wasn't a hesitation on the students end to attempt to reform the bot's responses when they didn't match expectations or the correct answer. This may be one of the largest benefits of using a chatbot as a proxy for supplemental instructors, being that there is no judgement, fear, or tact necessary in correcting a large language model when comprehension is on the line (although one student in particular found it quite pertinent to engage politely with Newt). Asking questions of a computer is part and parcel for generations who have had access to search engines for the entirety of their lives.

DISCUSSION

Many useful pieces of information can be extracted from this study. Although we were unable to find a statistically significant difference in academic markers, much is evident about the nature of engaging students in an academic environment outside of a student's natural interests. The design of this study has centered around the motif of creating engaging content that will invite students to participate. The question, it has become clear, is now *interesting to whom?* While Dr. Parker is right to note that learning is inherently a discovery, that discovery must also be something an individual is motivated to participate in [18]. People going to Disneyland do indeed learn while they are there, but as it turns out, there are some people who don't like going to Disneyland in the first place.

The lack of student participation in this study is a cautionary tale for those seeking to engage students outside of their natural context. This population was intentionally selected as a group meriting study because of the stigma associated with a learning environment that is a general education course. Otero may be pleased to find that we have eliminated one more microstate from the possible ideal configurations for classroom reform [7].

Furthermore, our findings seem to corroborate Cardinot's finding that learning is not a passive experience [19]. The high level of student queries lifted verbatim for analysis by Newt was surprising, but particularly in the fact that there was no long term benefit *or* detriment to student scores. This lack of an effect is likely due more to the limitations of the type of engagement rather than the resource itself.

In regard to the AI resource, this study has further illustrated that significant developments will be required for this technology to become a useful classroom tool. Indeed, as

Abdelghani so clearly pointed out, "…negative effects are due in particular to the lack of a pedagogical stance in these models' behaviors." [26] In order for a chatbot to serve as a digital tutor in any meaningful way, it is very probable that these assistants will need to be cut out from whole cloth rather than starting with the worldview baked into large language models which inundates users with information and asks for nothing in return.

Regardless of the resource, the literature clearly recognizes our need to meet students where they are at so that their development can continue unhindered [1-2, 16]. The students themselves clearly recognized that this pedagogical design was being employed, but more regularly chose to allow Newt to do the solving rather than asking the bot for assistance in understanding a concept.

CONCLUSIONS

This study has two primary questions. First, did the resources have a measurable impact upon student performance? Second, did the resources affect student perception of the sciences? To the first, we note that no statistically significant impact was had upon student scores. To the second, we note that the distribution of student's perspectives did fall more favorably in the end towards physics than it had previously.

Finally, although class engagement was not to the degree we had initially targeted, the lack of student participation is in and of itself a clear piece of evidence that many students likely are not very internally motivated to participate in class. The opportunity to receive additional educational materials was only accepted by 9.3% of the students enrolled in the course after three distinct, incentivized attempts to encourage students to participate in the study. For whatever various reasons students had, 90.7% of the class did not perceive a worthwhile the benefit-to-cost ratio for receiving free materials to support their learning.

These findings reflect a larger problem in classroom engagement. Quantitative analysis of the results necessarily degrades over time with the decrease in sample respondents. This is an obvious statement. What is less obvious—and merits further study—are pedagogical mechanisms that close the gap between educational practices and student interest. This problem is well known and documented in the literature, but unfortunately still remains unsolved.

Further research could be conducted on a class sample where the individual impact of each resource was tested in separate groups. This was the original plan for this study, but due to limitations in sample size was not feasible. Additionally, conducting a study with a larger sample size would prove beneficial. On a larger scale, further research could focus on how to motivate students to engage with these materials. These materials may have great potential in making an impact on student achievement, but the impact of any resource is inherently limited by the usage of its consumer. No tool is effective if it stays inside of its shrink-wrapped box.

CONFLICT OF INTEREST

No known conflict of interest exists for any of the researchers involved. This study was conducted without any benefit to any business, institution, or person with the sole exception of serving as a project in partial fulfilment of a Bachelor's of Science in Physics from Brigham Young University for the author of this study.

ETHICS APPROVAL

This study was approved by the internal review board of Brigham Young University on September 12th, 2024. A copy of the IRB approval can be found in Appendix B.

DATA AVAILABILITY

Data presented within the paper is publicly available for study, analysis, and critique. The raw data associated with this study is not publicly available in accordance with the BYU IRB guidelines and will be destroyed in accordance with federal regulation three years after the conclusion of the study to protect student anonymity.

ACKNOWLEDGEMENTS

Special thanks to Dr. Adam Bennion his continuous support, feedback, and comments to assist in the revision of this paper and the study behind it. Additional thanks to Dr. Jeannette Lawler for her critical analysis of this study. Portions of this study were funded by the BYU department of Physics and Astronomy.

As is always the case, thanks to Rebecca and Daisy who have make every day brighter.

APPENDICIES

APPENDIX A – EXCERPTS FROM PHSYSICS

The following sections are separated to present a section of the sample textbook *Phsysics* alongside a more traditional style and tone of textbook writing. Note that the "translations" were generated by ChatGPT. *Phsysics* is on the left, while the generated content is on the right.

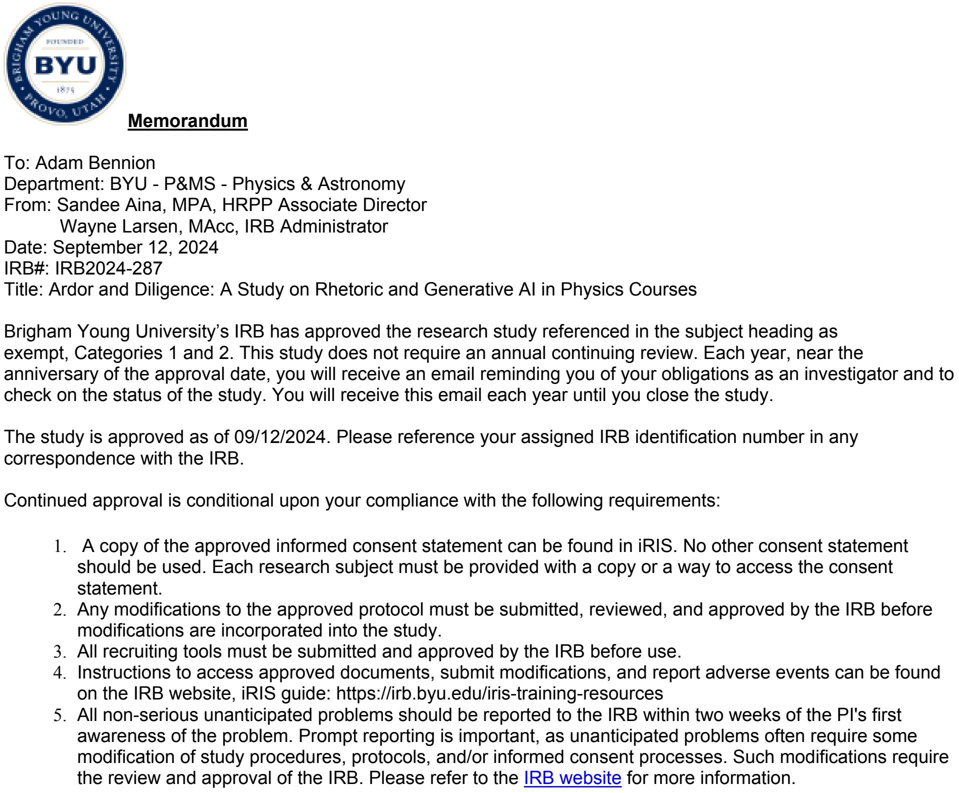
***Phsysics* Excerpt**	**Traditional Translation**
The Derivative: The Smallest Average in the World	*The Derivative: An Examination of Change*

A derivative examines how a mathematical relationship changes as it progresses. It is a numerical representation of how intensely a relationship is changing with time; the larger the derivative, the faster things are changing. Understanding how things change with time (or space!) is a very important tool: it allows us to effectively predict where a relationship will wind up in the future without having to wait around for it to happen. The derivative exploits the tool of division by looking at very, very small changes in a relationship. Effectively, a derivative is nothing more than an average on steroids. If you have seen what an average is before, you know that it is found through a very specific process: first, all of the terms are combined; second, the sum is divided by the number of terms in the set. Fairly straightforward. Doing this provides you with the mean: the value in- between all of the other values, with greater influence placed upon those values which show up more frequently in the set. The derivative differs from this basic method in three small yet meaningful ways: 1. There are only two values in a “set” of terms 2. The sum is replaced with subtraction 3. The spacing between the values is really, really, ridiculously small, but it matters	The derivative is a fundamental concept in mathematics that analyzes how a given relationship evolves over time. It serves as a numerical indicator of the rate at which a relationship changes; a larger derivative signifies a more rapid change. Understanding how relationships vary with respect to time or space is crucial, as it enables us to predict future behavior without waiting for the outcomes to unfold. The derivative utilizes the principle of division by investigating infinitesimally small changes within a relationship. Essentially, a derivative can be thought of as a refined version of an average, providing deeper insights into rates of change. To understand the derivative, it is useful to recall how a standard average is computed. The process involves two main steps: first, summing all the terms in a dataset; second, dividing that sum by the total number of terms. This straightforward procedure yields the mean value, which represents a central tendency, giving greater weight to more frequently occurring values. The derivative, however, diverges from this conventional approach in three significant ways: 1. It considers only two values in its “set” of terms. 2. The operation performed is subtraction rather than addition. 3. The distance between these two values is extremely small, which is crucial to its function. Through these distinctions, the derivative provides a powerful tool for analyzing and predicting change in various mathematical contexts.
Division	*Division*

Division is an operation that determines how many whole pieces of an object fit within a given space. It is useful for making *comparisons* about units and numbers. Think Tetris. Or perhaps you would like to know how many trees fit within a square piece of land in the forest. In the end, you would wind up with a *ratio* that says there are some number of trees *per* square piece of land. The operation is no different in Physics. The math involves taking one whole quantity and seeing what pieces fit. This process of finding a ratio between two things can be done with any two units you'd like. Velocity is a ratio between *distance* (meters) and *time* (seconds), which, not surprisingly, is measured in meters *per* second. Velocity gives you a single second to work with and asks, *how far can you get?* The bigger the number, the farther out you have traveled within your allotted time span of one second, and thus the faster you are going.	Division is a mathematical operation that determines how many whole units of one quantity fit into another. This process is essential for making comparisons between different units and numbers. For example, consider the game Tetris: understanding how pieces fit together mirrors the concept of division. Similarly, if you want to calculate how many trees can be planted in a specific area of land, division allows you to express this as a ratio, indicating the number of trees per square unit of land. In the field of physics, division serves a comparable purpose. The operation involves taking a total quantity and assessing how many smaller units can be derived from it. This ratio can be calculated between any two units. A prime example is velocity, which represents the ratio of distance (measured in meters) to time (measured in seconds), resulting in a unit of meters per second. Velocity effectively communicates how far one can travel in one second. The greater the velocity, the further the distance covered within that one-second interval, indicating a higher speed.
The definition of force is likely one you have seen before, though there are other, equally valid ways, of stating it. The most common expression to define force is the one below: $F = ma$ In English, this amounts to saying that an external force can be calculated by multiplying the mass of the system by the acceleration it experiences. This is what force is in our system. Other things might happen in the system, but unless it meets this condition (or an equal one) it isn't a force. A force is not, cannot, and will never be a mass times a position or a mass times a velocity, or anything other than exactly what amounts to a mass multiplied by an acceleration.	The concept of force is one that you may have encountered previously, although there are various valid ways to articulate it. The most widely recognized expression for defining force is represented by the equation: $\mathrm{F = ma}$ In this equation, "F" stands for force, "m" represents mass, and "a" signifies acceleration. In simpler terms, this equation states that an external force can be calculated by multiplying the mass of an object by the acceleration it experiences. This definition encapsulates the essence of force within our physical framework. While other interactions may occur within a system, they do not qualify as forces unless they adhere to this specific relationship.

When you encounter definitions in Physics, you have to be very careful to understand that what you are being told explicitly means one thing and one thing only. Understanding this early on will prevent a lot of misunderstanding and difficulty in answering physics problems.	It is important to note that a force is strictly defined as the product of mass and acceleration. It cannot be represented as mass multiplied by position, velocity, or any other variable outside this precise formulation. When engaging with definitions in physics, it is crucial to recognize that each term carries a specific meaning, which should not be conflated with other concepts. Developing a clear understanding of these definitions early in your studies will help mitigate misunderstandings and enhance your ability to solve physics problems effectively.

APPENDIX B – ETHICS APPROVAL



Memorandum

To: Adam Bennion
Department: BYU - P&MS - Physics & Astronomy
From: Sandee Aina, MPA, HRPP Associate Director
Wayne Larsen, MAcc, IRB Administrator
Date: September 12, 2024
IRB#: IRB2024-287
Title: Ardor and Diligence: A Study on Rhetoric and Generative AI in Physics Courses

Brigham Young University's IRB has approved the research study referenced in the subject heading as exempt, Categories 1 and 2. This study does not require an annual continuing review. Each year, near the anniversary of the approval date, you will receive an email reminding you of your obligations as an investigator and to check on the status of the study. You will receive this email each year until you close the study.

The study is approved as of 09/12/2024. Please reference your assigned IRB identification number in any correspondence with the IRB.

Continued approval is conditional upon your compliance with the following requirements:

1. A copy of the approved informed consent statement can be found in iRIS. No other consent statement should be used. Each research subject must be provided with a copy or a way to access the consent statement.
2. Any modifications to the approved protocol must be submitted, reviewed, and approved by the IRB before modifications are incorporated into the study.
3. All recruiting tools must be submitted and approved by the IRB before use.
4. Instructions to access approved documents, submit modifications, and report adverse events can be found on the IRB website, iRIS guide: https://irb.byu.edu/iris-training-resources
5. All non-serious unanticipated problems should be reported to the IRB within two weeks of the PI's first awareness of the problem. Prompt reporting is important, as unanticipated problems often require some modification of study procedures, protocols, and/or informed consent processes. Such modifications require the review and approval of the IRB. Please refer to the IRB website for more information.

REFERENCES (AIP)

1. C. S. Huai and W. W. Oo, "The influence of mathematical terminology on students' achievement at the high school level," J. Myanmar Acad. Arts Sci. 18, 9C (2020).
2. E. Schoerning, "The effect of plain-English vocabulary on student achievement and classroom culture in college science instruction," Int. J. Sci. Math. Educ. 12, 307–327 (2014), https://doi.org/10.1007/s10763-013-9398-8.
3. H. T. Williams, "Semantics in teaching introductory physics," Am. J. Phys. 67, 670–680 (1999), https://doi.org/10.1119/1.19351.
4. P. G. Hewitt, "The joy of teaching and writing conceptual physics," The Phys. Teacher 49, 412 (2011), https://doi.org/10.1119/1.3639147.
5. National Center for Science and Engineering Statistics, "Diversity and STEM: Women, minorities, and persons with disabilities 2023 (Special Report NSF 23-315)," National Science Foundation, Alexandria, VA (2023), https://ncses.nsf.gov/wmpd.
6. N. Murray, Writing Essays in English Language and Linguistics: Principles, Tips and Strategies for Undergraduates (Cambridge University Press, 2012), p. 147, ISBN 9780521111195.
7. V. K. Otero and D. E. Meltzer, "The past and future of physics education reform: The Every Student Succeeds Act, passed in 2015, harbors both threats and opportunities for physics education in the US," Phys. Today 70(5), 50–56 (2017), https://doi.org/10.1063/PT.3.3555.
8. P. Strube, "The notion of style in physics textbooks," J. Res. Sci. Teach. 26(4), 303–315 (1989), https://doi.org/10.1002/tea.3660260403.
9. I. YeckehZaare, P. Resnick, and B. Ericson, "A spaced, interleaved retrieval practice tool that is motivating and effective," in ICER '19: Proceedings of the 2019 ACM Conference

on International Computing Education Research, pp. 71–79 (2019), https://doi.org/10.1145/3291279.3339411.

10. J. Samani and S. C. Pan, "Interleaved practice enhances memory and problem-solving ability in undergraduate physics," npj Sci. Learn. 6, 32 (2021), https://doi.org/10.1038/s41539-021-00110-x
11. D. Rohrer, "Interleaving helps students distinguish among similar concepts," Educ. Psychol. Rev. 24, 355–367 (2012), https://doi.org/10.1007/s10648-012-9201-3.
12. V. X. Yan and F. Sana, "Does the interleaving effect extend to unrelated concepts? Learners' beliefs versus empirical evidence," J. Educ. Psychol. 113, 125–137 (2021), https://doi.org/10.1037/edu0000470.
13. T. Richter, L. Nemeth, R. Berger, R. B. Ferri, M. Hänze, and F. Lipowsky, "Using interleaving to promote inductive learning in educational contexts: Promises and challenges," Z. Für Entwicklungspsychol. Pädagog. Psychol. 54, 164–175 (2022), https://doi.org/10.1026/0049-8637/a000260.
14. D. Rohrer, R. F. Dedrick, and S. Stershic, "Interleaved practice improves mathematics learning," J. Educ. Psychol. 107(3), 900–908 (2015), https://doi.org/10.1037/edu0000001.
15. S. H. K. Kang, The Benefits of Interleaved Practice for Learning (Routledge, 2016), p. 12.
16. J. Kranda, "Precise mathematical language: Exploring the relationship between student vocabulary understanding and student achievement," Summative Projects for MA Degree, Math in the Middle Institute Partnership, University of Nebraska - Lincoln (2008), https://digitalcommons.unl.edu/mathmidsummative/7.

17. J. Donne, Meditation XVII, in Devotions upon Emergent Occasions (London, 1624).

18. J. Packer and R. Ballantyne, "Is educational leisure a contradiction in terms? Exploring the synergy of education and entertainment," Ann. Leisure Res. 7 (2004), https://doi.org/10.1080/11745398.2004.10600939.

19. A. Cardinot and J. A. Fairfield, "Game-based learning to engage students with physics and astronomy using a board game," Int. J. Game-Based Learn. 9(1), 16 (2019), https://doi.org/10.4018/IJGBL.2019010104.

20. V. Žák and P. Kolář, "Physics curriculum in upper secondary schools: What leading physicists want," Sci. Educ. 107, 677–712 (2023), https://doi.org/10.1002/sce.21785.

21. J. W. Schofield and others, "Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students," Soc. Sci. Comput. Rev. 8(1), 19–32 (1990), https://doi.org/10.1177/089443939000800104.

22. M. Raja and G. G. Lakshmi Priya, "Conceptual origins, technological advancements, and impacts of using virtual reality technology in education," Webology 18(2), 116 (2021), http://doi.org/10.14704/WEB/V18I2/WEB18311.

23. R. Winkler and J. Roos, "Bringing AI into the classroom: Designing smart personal assistants as learning tutors," in ICIS 2019 Proceedings (2019), https://aisel.aisnet.org/icis2019/learning_environ/learning_environ/10.

24. Lakshmi. G., B. S., R. D. M., D. J., and S. N., "AI-powered digital classroom," *Proc. 2022 Int. Conf. Commun. Comput. Internet Things (IC3IoT)*, Chennai, India, 2022, pp. 1–6, https://doi:10.1109/IC3IOT53935.2022.9767944.

25. R. Trisoni, I. Ardiani, S. Herawati, A. Mudinillah, R. Maimori, A. Khairat, D. David, and N. Nazliati, "The effect of artificial intelligence in improving student achievement in high

schools," in Advances in Social Science, Education and Humanities Research: Proceedings of the International Conference on Social Science and Education (ICoeSSE 2023) (2023), https://doi.org/10.2991/978-2-38476-142-5_50.

26. R. Abdelghani, H. Sauzéon, and P.-Y. Oudeyer, "Generative AI in the classroom: Can students remain active learners?" arXiv:2310.03192 [cs.CY] (2023), https://doi.org/10.48550/arXiv.2310.03192.
27. Bell, E. T. *Mathematics: Queen and Servant of Science*. p. 14.
28. Munroe, R. (2014). *What if? : Serious scientific answers to absurd hypothetical questions*. Houghton Mifflin Harcourt.
29. Kakalios, J. (2005). *The physics of superheroes* (2nd ed.). Gotham Books.
30. W. Shakespeare, "Brevity is the soul of wit," in Hamlet, Act 2, Scene 2 (ca. 1602).

[i] This is a significant consensus in and of itself and deserves attention for further application in the classroom. Many studies have been conducted in this particular vein and indeed this study could serve as a launching point on the practicality of implementing these ideas in future research.

[ii] Whether or not Newt “chose” to adhere to these guidelines is another story. Once again, a potential trove of studies could be conducted on the willingness of large language models to implement their instructions with precision. Although this is significant, we will not take much time to address it as it is not the focus of this particular study.