UNIVERSITY OF CALIFORNIA
Los Angeles

Applications of Large Language Models in Education: Literature Review and Case Study

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

John DuChateau Baierl

2023

ABSTRACT OF THE THESIS

Applications of Large Language
Models in Education: Literature
Review and Case Study

by

John DuChateau Baierl
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Mark S. Handcock, Chair

The rapid rate of improvement of natural language processing (NLP) systems and large language models (LLMs) begets a wide array of applications in the field of education and classroom instruction. The possibility of individualized practice and immediate student feedback from a low-cost and widely-available service has an enormous capacity to change modes of student instruction. In this review, we discuss the current state of research into the applications of LLMs for science and mathematics classroom education, calling particular attention to concerns surrounding overreliance and equity, as well as suggesting specific directions for future study. We conclude by considering the *CourseKata* interactive textbook as an illustration of how AI tools may begin to reshape traditional methods of content delivery.

The thesis of John DuChateau Baierl is approved.

James W. Stigler

Qing Zhou

Robert L. Gould

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2023

To my family and

Lindsay, for putting up with

my nonsense

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the UCLA Teaching and Learning Lab and my colleagues in the Department of Statistics for support and guidance.

CHAPTER 1

Introduction

Large language models (LLMs) such as ChatGPT and BERT [9] have grown dramatically in popularity for both academic and personal use. ChatGPT in particular is estimated to be the fastest-growing consumer application in history [21]. Developments in transformer architectures and self-attention mechanisms [43] have enabled these models to better handle long-range dependencies in text and produce coherent and useful conversational responses. Pre-trained on an extensive corpus, these models demonstrates state-of-the-art performance at a remarkable variety of tasks ranging from essay-writing to generating quiz questions on desired topics.

As organizations in many sectors, including schools, have already begun to implement regulations and occasionally outright bans surrounding the use of ChatGPT in academic work, a deeper discussion on the applicability of artificial intelligence (AI) tools in education is inevitable. Systems like ChatGPT undoubtedly hold enormous appeal as teaching tools. Their potential to enhance student interactivity, provide tailored feedback, and create personalized learning materials make LLMs a viable means of addressing an array of classroom needs.

Moreover, as increasing class sizes in United States public schools continues to receive substantial focus [32], the democratization of high-quality and personalized math tutorial holds great potential to close gaps in access to resources along socioeconomic lines. The recent growth in the popularity of private tutoring among disproportionately wealthy private school students suggests that such services are both in high demand and also highly unevenly distributed among students at varying levels of wealth, with some tutors and agencies providing personalized academic guidance charging rates upwards of $400 per hour [2]. Providing a similar quality of widely-available supplementary individual instruction at low or no cost holds enormous appeal given this inequity.

However, as AI systems begin to work their way into a wider range of economic sectors it is increasingly important that the actual capabilities of generative AI systems be accurately represented. As we will discuss, this is especially true in the context of science, technology, engineering, and mathematics (STEM) education. An accurate account of the capabilities, weaknesses, and biases of LLMs is a prerequisite for developing effective AI implementations for student use that exploit its strengths without exacerbating existing current issues in STEM classrooms.

The primary objective of this paper is to outline the present body of research on applications of large language models and generative AI systems in general for both K-12 and postsecondary classroom use, with a particular focus on mathematics education. We spotlight uses of generative AI in creating assessments for student use and consider the potential for open interaction between learners and automated systems. We further seek to highlight their specific virtues and shortcomings that should inform their role in the classroom. While the primary focus will be on science and mathematics education, many issues raised are broadly relevant across disciplines. We conclude by discussing the *CourseKata* interactive textbook developed through the UCLA Teaching and Learning Lab as a case study for understanding these concerns in a concrete context.

CHAPTER 2

Literature Review

2.1 Uses Generative of AI in Education

The idea of utilizing AI systems for educational tasks is not a recent one. Discussions of algorithmically generated learning materials date back to the 1970s [11]. However, the rapid growth of generative AI in the fields of natural language processing (NLP) and computer vision have opened a wide array of uses both as a tool for in-class and supplementary instruction, accelerating the discourse in the recent years.

An emerging body of research has investigated the effectiveness of such tools in the classroom and demonstrate their enormous potential for automatic question generation and direct interaction with LLMs. Prior reviews primarily focused on outlining potential applications of LLMs in education and highlighting the need for additional literacy among both students and educators to better understand the technology, such as Kasneki et al (2023) [25]. The authors highlight future concerns such as the potential for student over-reliance on models to erode critical-thinking and problem-solving skills. These are important considerations should indeed guide specific implementations of LLMs and algorithmically-generated content into learning materials.

However, less attention has been paid to the ways in which such tools may fit within present student-teacher dynamics, and the degree to which algorithmically-generated course content is likely to reshape the role of educators and teaching content. Moreover, while critical issues like encoded biases in NLP systems have been well-documented [5], less attention has been paid to how AI shifting the role of instructors interacts with present inequities in STEM and how this brings general trust in AI systems to bear. While much of this research is still in relative infancy with limited empirical study [22], we give a brief outline of work done to date on applying AI and NLP systems in education, as well as directions of continuing research.

The growing body of work in this field has found generally positive results in the ability of LLMs to produce useful learning materials and serve as fruitful conversational agents with learners [37] [23]. A significant virtue of incorporating instruction via interaction is that such tools better incorporate elements of personalized interaction to otherwise remote learning activities. This allows for striking what Vie et al (2017) describe as "a better balance between giving learners what they need to learn (i.e. adaptivity) and giving them what they want to learn (i.e. adaptability)." [57] In short, NLP tools like GPT-3 and its relatives help to alleviate the top-down nature of traditional approaches to remote student work.

Incorporating open-ended conversations and responses to prompts generated by chatbots is one such application toward this end that has received substantial study. Steuer et al (2021) found automatically generated questions to be relevant to their intended topics, free of language errors, and to contain natural and easily-comprehensible language in a variety of domains using their autoregressive language model [52]. Additionally, their generated questions successfully addressed central concepts of their training texts and topics, which the authors describe as pedagogical "coreness". This suggests that the produced tasks were indeed pedagogically useful within their subjects and contexts.

While this is an encouraging result, the notion of pedagogical coreness is difficult to pin down and is highly subject- and instructor-dependent. For instance, one middle school astronomy curriculum might emphasize storytelling from limited data as a recurring theme, while another might focus on spatial reasoning in three dimensions. Of note is that Steuer et al limited their study to automatic question generation from individual textbooks, relying on expert judgement to assess the degree to which the questions produced aligned with central information from the text.

Though useful questions are essential, assessing how students respond to and interact with them is also needed. Abdelghani et al (2023) compared question-asking behavior among primary school students after utilizing the prompt-based learning of GPT-3 to directly automate elements of course content [1]. This was a particularly encouraging result, since it featured a more open-ended interaction structure and a greater focus on student responses than Steuer er al (2021), giving some indication of how prior results might generalize to LLMs applied to an even wider range of possible tasks. Overall, their results suggest that such automated prompts generally elicited positive responses from students and show potential for increasing curiosity and feelings of agency in their learning.

Additionally, Wu et al (2020) found that interaction with a chatbot in E-learning environments alleviates feelings of isolation and detachment that often accompany the use of such platforms [60]. As more learning content has shifted online in the wake of the COVID-19 pandemic [4], being able to provide access to high-quality instruction regardless of time and place is relevant both for present and future impact on pedagogy.

While these results present highly encouraging paths forward, it is important to consider the limited scope of much of the research conducted to date. Though the studies discussed above involved some degree of open-ended interaction, they were largely limited to providing prompts or keywords within a narrow task framework. Fully open-ended chatbot-style conversations for pedagogical uses has yet to receive specific attention.

A natural question in this setting is the degree to which information and reasoning provided by AI agents is reliable. Jiang et al (2021) investigated the calibration of LLMs on trivia-style question-answering tasks across a number of disciplines [23]. They found that while the models tested (GPT-2, T5, BART) performed well, they were generally poorly calibrated, tending to be over-confident in their predictions. The authors show that model fine-tuning procedures substantially improve this issue. While less critical than fields like medical diagnosis where safety and proper confidence calibration are essential, properly calibrated degrees of confidence are highly relevant for student feedback and interactions. Future work should build upon the degree to which domain- and class-specific fine-tuning can improve LLM reliability at question-answering tasks within a student-AI interaction setting.

In addition to creating course content and engaging students in discussions, developments in generalized adversarial network (GAN) architectures [18] and AI-generated media allow for systems that produce synthetic interfaces with which students can interact. Pataranutaporn et al (2021) discuss potential use cases of AI-generated animated characters utilizing GAN architectures for interaction in learning environments [45]. Prior research demonstrates that learning materials incorporating interaction with fictional characters positively impacts student experiences, improving motivation and attitudes [30].

This work suggests a strong potential for generative AI to enhance both the instructional content being delivery, but also the mode of delivery itself in ways the promote motivated and curious engagement from students across age and ability spectra. This is a particularly intriguing area of research, since these early results align well with the findings of We et al (2020) of AI-interaction reducing some of the prominent downsides of loneliness in online learning environments for students. Future work should seek to combine LLM interactions with GAN-created animations, allowing interactive learning content to be enjoyable and highly interactive for younger students as well.

2.2 Unique Challenges Associated with STEM Education

The application of language models to STEM education presents a uniquely challenging use case for LLMs in a variety of ways. Quality mathematics instruction requires a wide array of skills such as acute social awareness, deep understanding of course material, a clear sense of the long-term skills required for later work in the field, a sense of aesthetics and student interest, as well as an understanding of precisely why some topics are challenging or counterintuitive. Some these skills are well-suited to collaboration with AI systems, while others have proven to be more challenging. In this section, we focus on the aspects of STEM education that make applications of NLP uniquely difficult and highlight where they might offer significant value.

Teacher-student relationships have been shown to have significant impacts on students' perceptions and feelings toward STEM fields [39] [7]. Similarly, Skinner and Belmont (1993) find that teacher involvement is central to positive classroom experiences for young students [50]. However, they also highlight a feedback effect of student motivation on teacher behavior; students demonstrating disengaging behavior were more likely to receive responses from instructors that further undermined their motivation. Considering how LLM tools fit into this setting, a highly desirable feature its insulation from perceptions of student motivation.

Here we see both the potential and pitfalls of a naïve implementation of AI systems in the classroom. While such tools are able to ameliorate instructional challenges caused by time constrains or implicit bias toward perceived student motivation, they also run the risk of displacing the desirable outcomes from a productive student-teacher relationship. Many of these issues are even more pronounced in mathematics classrooms. LLMs rely on the assumption that the choice of words used accurately reflects the desired concepts. In other words, that the embeddings resulting from the pre-training of the model are properly specified for an individual's use of language in the prompt. Bender and Koller (2020) emphasize this, stressing the need for the distinction between linguistic form (any observable realization of language: words, symbols, etc), and meaning (the relation of the form to some concept external to language) when discussing the capabilities of LLMs [6]. Fundamentally, language models are trained on respond to linguistic form with limited understanding of the degree to which they capture meaning.

The application of LLMs to a teaching context brings this issue to the forefront. A major focus of teaching in an introductory course is developing the alignment between mathematical "form" and mathematical "meaning." Indeed, vocabulary acquisition and usage in math learning is notably unreliable among students [40]. A central task of math educators is identifying and correcting these misconceptions to enable students to hold productive mathematical conversations. This development of fluency in the language of mathematics is a critical step that enables deep growth in the field. However, the nature of LLM training makes these models uniquely ill-suited to these tasks, where diagnosing misalignment between intent of the speaker the specific language that they use is paramount.

One approach to addressing this issue is directly investigating LLMs' robustness to language mis-specification. While the robustness of ChatGPT has been extensively studied [59], this has typically been done through a lens of AI safety. Less work has emphasized the specific types of mis-statements or vocabulary errors common in mathematics classrooms. Such insight is needed to inform the specific teaching tasks that LLMs are best equipped for, and what types of interventions and clarifications are better-suited to human instruction.

While prior reviews such as Kasneki et al (2023) rightly emphasize the potential for both student and instructor over-reliance on interactive LLM instruction, it is worth further articulating this point in the specific context of math instruction. Teaching math effectively requires a delicate balance between over- and under-explanation, relying on a variety of cues of when to intervene, and when to allow students to struggle and create on their own [29].

A useful illustration of this can be found in Paul Lockhart's *A Mathematician's Lament* [38]. Lockhart heavily criticizes American mathematics education, focusing on the rote, shallow, and joyless nature of math problems found in many textbooks and math courses. Over the course of *A Mathematician's Lament* (and his subsequent *Measurement*), Lockhart outlines how the conception of the math problem can be re-thought to better encourage students to deeply engage in mathematics as a practice, rather than merely learning mostly-rote mathematical facts as trivia. He describes the following geometry problem assigned to his middle school math class as an illustration of this idea.

The problem is to explain the surprising property that any triangle placed in a semicircle as shown below forms a right angle, regardless of where we place the tip of the triangle. This is a simple yet surprisingly deep investigation that exercises an aspiring mathemati-

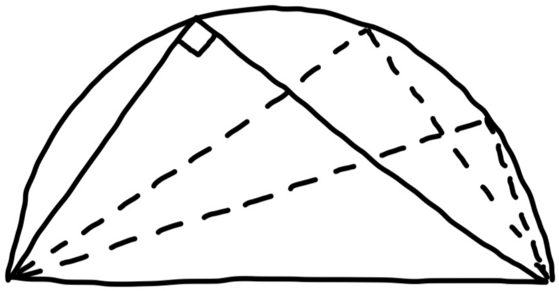

Figure 2.1: Student sketches: triangle in a semicircle

cian's ingenuity. The student produced the following argument (rephrased by Lockhart for publication):

"Take the triangle and rotate it around so it makes a four-sided box inside the circle. Since the sides of the box must be parallel, so it makes a parallelogram. But it can't be a slanted box because both of its diagonals are diameters of the circle, so they're equal, which means it must be an actual rectangle. That's why the corner is always a right angle."

This student can carry a number of deep mathematical lessons from the experience of working on this problem:

- The ability to tinker and play when tackling a novel, challenging problem
- The usefulness of symmetry
- Understanding of the role that mathematical drawing plays in developing the feel and intuition of a problem
- The appeal of parsimony, beauty, and aesthetics of mathematical argument—this is a much more elegant proof than the one found in many geometry textbooks!
- Seeing a novel argument emerge from a series of failures

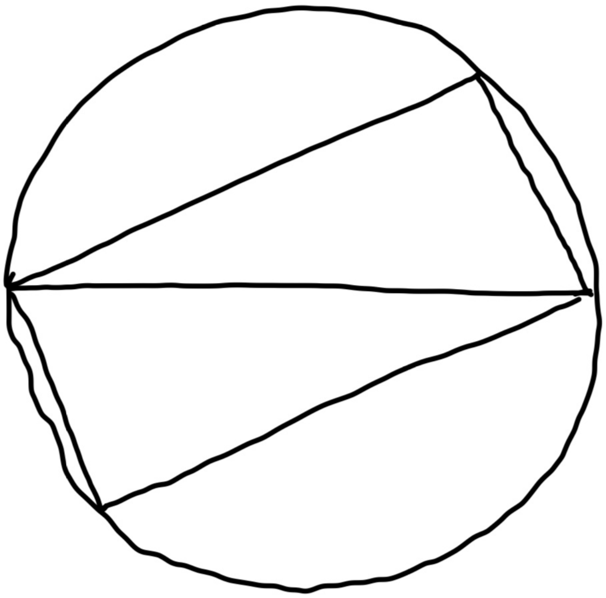

Figure 2.2: Student solution: flip it and reverse it

These are high-level, long-term objectives for mathematical learning. Identifying moments in which students' questions, comments, or points of confusion present opportunities to gradually build these high-level and abstract skills requires a keen awareness of not only the core material of a chapter or even an entire course, but also of the broader landscape of the subject of math. Crucially, the instructor must recognize habits in students' thinking can be readjusted and developed further to build both long-term successes, but also long-term enjoyment of the subject.

The specific words that instructors choose to use are based both on knowledge of the core subject matter, which research suggest that AI-generated feedback mimics reasonably well, but also how to actively curate moments that build these high-level skills. Instructors make decisions such as when to intervene, when to probe, and crucially, when to simply say nothing and allow the student to explore on their own. It is our position that these choices require an understanding of the feelings of frustration and writer's block that accompany such work in addition to a clear grasp of mathematical content. The instructor is facilitating and guiding a student through the experience of doing math as much as delivering definitions and content, and perhaps moreso.

An overreliance from both the student and instructor on algorithmically-generated advice risks displacing these essential experiences in learning mathematics. While prior research affirmed the relevance and quality of automatically generated questions with respect to course material, less attention was paid to the degree of relevance with respect to building the type of general mathematical skill outlined above. The degree to which NLP systems are able to capture these elements, and the optimal way to make full use of their utility while enabling human instructors to fill in these gaps are critical questions that deserve further investigation.

Moreover, no consideration of the importance of selectively withholding guidance as part of the learning process in the AI-for-education literature was found in this review. However, this is an essential design consideration to avoid the student-level overreliance discussed to date. Research into the use of LLMs to build student question-asking and build student curiosity are useful jumping-off points for continuing this line of research.

Equally important is that instructors not over-rely on algorithmically generated questions that, while they may be well-formed and relevant to core course material, lack unity with respect to core mathematical thinking skills and variety or aesthetic appeal with regards to their reflection of those. This role of mathematics instruction as facilitating experiences that reflect what it means to “do” math should inform how AI systems are incorporated into its teaching. Though this specific issue has received limited study, some parallels can be drawn from prior work surrounding the experiences of students in computer-based instruction in general. Krupa et al (2014) found that while students in computer-based Intermediate Algebra outperformed those in a parallel face-to-face section, they showed more limited ability to interpret equations and relate them to concrete situations [33].

With these issues in mind, balancing face-to-face instruction and curriculum design with AI augmentation should seek to use LLMs as a more engaging platform for developing necessary rote mathematical and technical skills through more valuable and personalized practice, setting up human teachers to dedicate more class time to emphasizing applicability and holistic understanding of progression through the field. Evidence suggests that well-implemented LLMs are highly capable of delivering well-phrased mathematics content and developing student curiosity for the subjects, which teachers can then capitalize on.

2.3 Trust, Perceptions of Fairness, and Algorithmic Awareness

Trust in educators and developing mentorship relationships are critical for students' long-term success in STEM fields [31]. As teaching tasks are shifted from human instructors to LLMs, perceptions and trust in those algorithms is an important consideration when thinking about how those relationships will respond to such a transition. This is a particular concern for user trust in the output of LLMs, the decisions of which are typically a black box for both users and implementers.

Literature about general trust in AI and perceived fairness of algorithmic decision-making provide a useful starting point for understanding how reliance automated systems might map onto STEM classroom settings. Extensive research has examined perceptions of algorithmically-generated decisions in contexts ranging from Facebook news feed curation to medical diagnosis [36] [27]. Eslami et al (2015) examined how providing users explicit information about the presence and functionality of decision-making algorithms affects the user experience [14]. They measured the effects of revealing the difference between algorithmically-curated and unadulterated Facebook News Feed content to users, finding that awareness generally increased the level of satisfaction with the platform. This suggests that awareness of the presence and functionality of recommender algorithms is a meaningful consideration for the user experience.

However, in an NLP-for-education context, there is no equivalent of an unadulterated feed as in Eslami et al (2015). The more important comparison in a teaching setting is between decisions made by an algorithm and ones made by humans rather than some neutral baseline. Lee (2018) did just this, more directly comparing attitudes toward decisions in a managerial scenario made by human with those made by AI system [35]. They found that for tasks involving primarily mechanical skills, algorithmically- and human-made decisions were perceived as equally fair. The authors adopt a definition of fairness as ”treating everyone equally or equitably based on people’s performance or needs,” emphasizing perceived fairness rather than algorithmic fairness in their work. Crucially, for decisions requiring human skills such as hiring or work evaluation, algorithmic decisions were rated as less fair by this criterion.

Kizilcec (2016) further investigated the relationship between different levels of algorithmic transparency in the user interface and perceived fairness in the context of work evaluation [28]. They found that users whose expectations matched the algorithm’s feedback trusted those decisions regardless of the level of transparency provided. However, trust levels of the users who received a lower rating than expected depended on the level of transparency provided. While some degree of explanation improved attitudes, too much transparency again eroded perceived trust.

This suggests that the way in which AI systems are presented to students and the amount and type of background information they are provided about their functionality are likely to affect their experience and level of trust, particularly when applied to a high-stakes or highly-personal application like academic feedback. NLP algorithms are not yet able to accommodate social values such as fairness, nuance, or context [34], making these structural choices and transparency vital. Care should be given as to the degree to which students are made aware of the underlying model, explaining its functionality, and coached on how to use and interpret their outputs and instructions. This warrants specific exploration in classroom applications to explore the degree to which these sentiments translate across settings and context. See Dodge et al (2019) for further discussion of how algorithmic explanation impacts perceived fairness [12].

Another significant concern moving forward with implementations of AI and LLMs as teaching tools is that trust in algorithmic decisions is not constant across demographic groups [41]. This is potentially highly impactful for deploying AI systems in STEM, a field which already faces substantial issues of udnerrepresentation. Strong emphasis should be placed on more closely understanding how these effects interact with each other in education settings specifically. Literature on algorithmic trust from related fields suggests that specific details in how these tools are presented and incorporated in materials can have a strong effect on levels of trust, potentially mitigate some of these undesirable effects. Moreover, the presence of a human instructor may temper this distrust to a degree. Kricorian et al (2020) emphasize the crucial role that human mentorship plays in bridging underrepresentation gaps [31]. One objective in AI implementation ought to be freeing up additional time for instructors to spend on the types of mentorship-style activities associated with higher engagement from students in underrepresented groups in STEM.

2.4 Equity in STEM

A major concern in implementing NLP systems in STEM education is the current state of underrepresentation along gender, racial, and socioeconomic lines in the many sectors of the field [24] [15]. Given the challenges associated with identifying the causes of that underrepresentation in select STEM fields, it is critical that researchers and implementers of AI systems strongly consider the potential effects of AI trust interacting with these existing inequities in STEM fields and education. We provide a brief overview here to highlight potential points of concern moving forward.

LLMs do bring some unique assets from an equity standpoint, such as their insulation from negatively responding to perceived student motivation as previously discussed. However, these systems also carry the potential for carrying their own biases reflected from their training corpus, with little possible recourse by due to the black box nature of their decision-making. Though these are certainly relevant to student-facing AI interfaces, the discussion ought to be expanded to address the extent to which a transition to algorithmically-generated instruction exacerbates existing issues associated with equity in STEM.

One drawback of replacing traditionally human-driven teaching tasks with AI interaction is the loss of the positive impacts that the relationship-building stemming from those conversations has on students. Kricorian et al (2020) suggest that students' decisions to pursue further STEM education and career tracks is highly driven by positive mentorship experiences from someone sharing their gender or ethnic identity [31]. Implementations of AI systems that seek to fully replace face-to-face, tutorial-style interactions run the risk of removing this driver of students from underrepresented groups in STEM into the field. On the other hand, an AI system that is implemented in a way that redistributes the role of the teacher in classrooms relying heavily on lecture in order to increase time spent in small-group or individual instruction has the potential to facilitate more of these mentorship-building student-instructor interactions.

AI use also presents a number of opportunities to specifically address strategies for closing underrepresentation gaps. Wang and Degol (2017) recommend a focus on developing interest in STEM subjects in addition to developing aptitude in order to close gender gaps in STEM among young students [58]. Additionally, digital storytelling strategies develop engagement and interest in the field [48] [26]. These approaches lend themselves well to the GAN-produced animations and previously-discussed methods that promote engagement and student-guided interaction.

2.5 Lessons from Adversarial Policy Literature

A frequent concern in media discourse surrounding applications of AI in industry is the extent to which these tools reshape or replace human labor in those positions [17] [56]. As research continues in the viability of LLMs for educational tasks, it feels inevitable that these conversations will shift to consider the possibility of AI displacing human educators entirely. The position of this paper is that such an arrangement would be sub-optimal and would lead to numerous undesirable consequences. However, given the seeming inevitability of such proposals, it is worth highlighting several high-level conceptual shortcomings of current state-of-the-art AI across multiple fields to gain some sense of how this conversation may evolve moving forward.

AI systems applied to logic games present a useful setting in which to consider their shortcomings. As environments with clearly defined goals, discretely defined piece positions, and complete deterministic control over piece movement, these games present an extremely friendly learning environment for AI reinforcement learning algorithms such as AlphaGo [49]. Nonetheless, evidence from the literature on adversarial attacks and policies suggest several overarching weaknesses in backpropagation-trained learning networks. Wang et al (2023) describe and test a conceptually-simple adversarial policy for the board game Go, beating the KataGo reinforcement learning algorithm sufficiently trained to superhuman performance levels in over 70% of attempts [59]. Their approach employs a highly suboptimal and conceptually-simple scheme that KataGo nevertheless struggles to counter.

It is important to note here that the concern stemming from these results is not merely that algorithms relying on pre-trained, deep learning architectures are fallible. The more meaningful observation is that these examples show a surprising inability to detect simple, high-level logical connections that are quite trivial for human reasoning. This is reflected in Richardson, Heck (2023) as well, finding that despite their remarkable capabilities, state-of-the-art NLP systems still struggle with even common-sense reasoning that humans find trivial. Niven, Kao (2019) also highlight that many situations in which LLMs appear to perform abstract reasoning can be reduced to the models leveraging statistical artifacts, which leave them vulnerable to adversarial attacks that reduce their performance to random guessing [42]. Talmor et al (2019) perform deeper analysis on the degree to which pre-training captures logical reasoning in a zero-shot setting [55]. They again highlight significantly varying performance across different models, with some models failing entirely on particular tasks. Recent work by Bubeck et al (2023) has suggested high-level reasoning capatilities of GPT-4, the successor to GPT-3 and ChatGPT [10]. However, GPT-4 still struggles to hold and follow conceptual mathematics discussions, despite its improved aptitudes for question-answering and problem-solving.

Teaching effectively requires precisely this type of lateral, high-level strategic thought. Go presents a kind environment for reinforcement learning performance, and in many ways is something close to a best-case-scenario for the applicability of learning algorithms to recreate abstract reasoning skills humans traditionally rely on to succeed. Educational instruction at a high level presents a much more wicked learning environment that requires selecting the task that needs to be performed (Socratic instruction, didactic explanation, or simply not intervening at), the mode in which content should be delivered (pictorially, verbally, in writing, or acted out) in addition to selecting the particular words or symbols to deliver to the student.

While this discussion provides only a loose analogue to the NLP-for-education setting, there are general parallels that can be drawn to inform areas for caution as teaching tasks become shifted toward automated systems pre-trained by backpropagation. Given the inability of AI systems across several fields to perform higher-level, abstract reasoning about concepts, we should expect that an optimal division of labor between human and AI instructors maintains some role for human instructors given the current technology. Moreover, this emphasizes the social value that instructors provide in developing deep interest for their subjects, and that a high-quality implementation from AI should seek to complement and highlight those roles.

CHAPTER 3

CourseKata Case Study

The *CourseKata* [51] online textbook presents a useful case-study of an interactive learning tool ripe for applications of AI, NLP and LLM tools. Statistics education sits at a unique nexus of theory and application. Traditional advanced high school and introductory undergraduate-level statistics courses (such as Advanced Placement Statistics) cover a large and seemingly disconnected body of content, spanning elementary probability theory, data visualization, classic summary statistics, linear modeling, and statistical testing.

The *CourseKata* project aims to address these challenges within traditional introductory statistic curriculum, emphasizing development of a more robust and unified understanding of statistical thinking. The curriculum emerges from what the authors refer to as the practicing connections framework [16], which suggests that effective instruction should facilitate connections between a domain's core concepts, key representations, and contexts and practices of the world. The project focuses on a view of learning that shifts emphasis away from "bits" of knowledge—independents facts and procedures—toward a more coherent and flexible connectedness of concepts that better represents expert fluency in the domain. The book is aimed at upper high school and undergraduate introductory statistics students and is available in multiple versions tailored to different course levels.

The *CourseKata* textbook poses questions to students in three primary ways. Items are embedded within chapter readings to develop understanding of course material, as well as at the end of sections within each chapter to reinforce topics. Lastly, summative assessments of approximately 15-20 questions are given at the end of each chapter to measure cumulative grasp of core chapter and course goals. These include multiple choice, multiple-select, and open-response items.

Frequent interactivity in the form of short coding tasks, multiple choice questions, and free-response questions within the text addresses the shortcomings of passive textbook reading as a learning tool, facilitating deeper understanding and more robust mental representations that generalize effectively to new tasks and related concepts. This is consistent with the extensive literature that posing questions about reading facilitates higher-quality and transferrable learning [47] [3].

CourseKata presents a useful opportunity to consider how a more standard approach to textbook question content and presentation might be re-thought with the considerations from our literature review in mind. LLMs show promise both as useful formative learning tools and in developing summative tasks for students, so AI implementations accompanying both measures of student progress within the current textbook versions are viable.

CHAPTER 4

Data

Anonymized response data is available directly from *CourseKata* for any registered course use the textbook. The necessary sample sizes to produce stable item-response parameter estimates limit our analysis to large class sections. Drasgow (1989) found that sample sizes of at least 200 produce stable parameter estimates using marginal maximum likelihood estimation for test instruments of at least 5 items [13]. The sample for this analysis was drawn from a class of 240 students at a large, highly-selective public university. 25 students from the course opted out of sharing their data, resulting in data available from 215 respondents.

The portion of *CorseKata* textbook used in our sample is split into chapters as follows:

1. Welcome to Statistics: A Modeling Approach
2. Understanding Data
3. Explaining Variation
4. Examining Distributions
5. A Simple Model
6. Quantifying Error
7. Adding An Explanatory Variable to the Model
8. Models With a Quantitative Explanatory Variable

Note that at the time of this writing, the most recently completed courses used version 3.0 of the textbook. The complete textbook and assessment questions considered below are freely available through the *CourseKata* website.

CHAPTER 5

Analysis and Discussion

5.1 Model-Fitting

There are several connections to be drawn from our prior review when considering how AI systems might be implemented in *CourseKata*. The chapter review assessments are one area of the textbook that might be rethought in light of the incorporation of AI-assisted learning. Given the encouraging research surrounding usefulness and quality of AI-generated questions, an intriguing path forward could be to replace some or all of these hand-written items with LLM-generated ones. This would allow for students to play a role in actively shaping their own review materials, allowing them to receive focused practice with content they personally find to be more challenging.

An important step in considering this change is to better understand the role that the current hand-written *CourseKata* assessments are playing in students' learning. To accomplish this, we turn to the item response theory (IRT) framework. Our goal here is to develop a clearer picture of the level of difficulty of these textbook test items with respect to the chapter content, and the degree to which they are providing useful and interpretable information to instructors about student progress. With the potential for some amount of future question-generation being ceded to black-box AI systems, tailoring the hand-selected task to maximize interpretability becomes all the more essential to monitor overall class progress and identify specific areas in which students may be struggling.

The two-parameter logistic (2PL) IRT model is a useful framework here, providing information both about the item difficulty and the item discrimination to indicate how effectively test items are locating student ability along the latent dimension. The model is shown below:

$$P(X_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \frac{\exp(\alpha_j(\theta_i - \delta_j))}{1 + \exp(\alpha_j(\theta_i - \delta_j))}$$

This predicts the probability of student i answering correctly on dichotomous item j, where δ_j and α_j denotes the item difficulty and item discrimination parameters and θ_i is the latent ability parameter for that student.

5.2 Dimensionality Assessment

Note that this is the one-dimensional 2PL model. This can be generalized to higher dimensions depending on the hypothesized underlying latent dimensionality of the item-response data. While IRT models tend to be somewhat robust to mis-specification of the underlying dimensionality, assessing the dimensional structure of a measurement instrument is a critical step when applying IRT models. Analysis of the performance of popular methods for dimensionality assessment in the MIRT context finds that traditional parallel analysis using principal component analysis and tetrachoric correlation performs best at achieving the highest proportion of identified underlying dimensions [19]. This held both when the generated IRT model is unidimensional and multidimensional.

We employ this approach on the Chapter 1 Review responses from our sample. Chapter 1 review consists of 14 total questions, all of which are multiple choice. For simplicity, we treat test items as dichotomous. Chapter 1 introduces the basics of R programming, with test items emphasizing utilizing functions and interpreting R syntax. Given this relative conceptual unity of this content, we hypothesize that a unidimensionality assumption in the latent space might be reasonable.

The scree plot is shown below of a parallel analysis with 1000 Monte Carlo iterations performed on the Chapter 1 item matrix. The results indicating that for only a single latent factor did the eigenvalue exceed that of the random data. This supports the hypothesis of unidimensionality, so we proceed under that assumption.

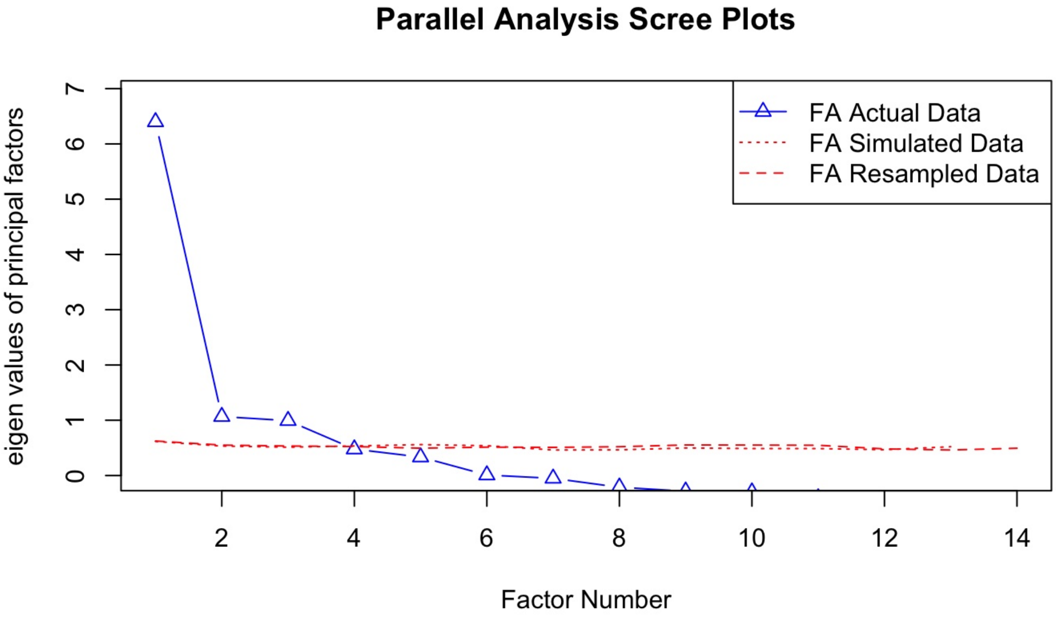

Figure 5.1: Scree plot

5.3 Model Analysis and Discussion

The full table of parameter values for the fitted model is available below. The most obvious trend here is the generally low values of the item difficulty parameters. These suggest that the primary value of this test instrument is providing practice at relatively straightforward skills. The test instrument as a whole provides information at a relatively low ability level along the single latent dimension, as shown by the full information function plotted below. Note that this is simply the sum of the individual item information functions.

The Wright map for this model provides further insight. The difficulty parameters for the

	α	δ		α	δ
Q1	6.756	-2.726	Q8	2.332	-2.517
Q2	1.840	0.870	Q9	3.308	-2.093
Q3	0.962	-3.607	Q10	2.312	-2.242
Q4	2.293	-3.034	Q11	2.135	-0.737
Q5	2.472	-2.742	Q12	2.623	-2.148
Q6	0.614	-2.508	Q13	1.199	-1.941
Q7	1.777	-1.854	Q14	1.245	-3.519

Table 5.1: Parameter values from fitted 2PL model for Chapter 1 review

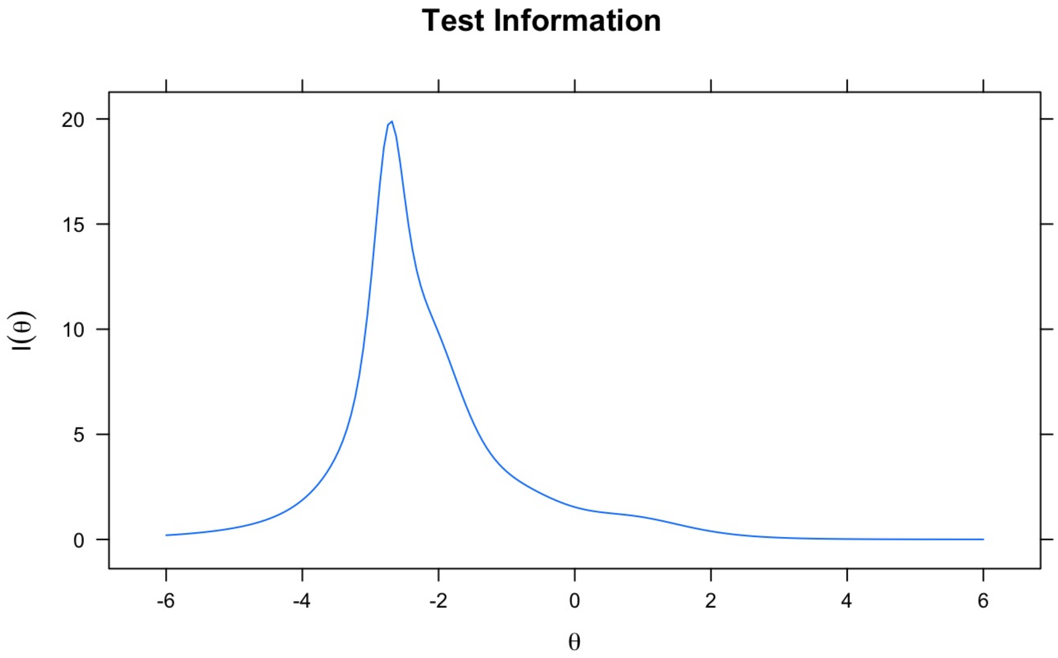

Figure 5.2: Chapter 1 Review: Full-test item information function

14 individual questions from the Chapter 1 review are plotted along the horizontal axis, with the item difficulty parameter δ_j plotted along the vertical axis. Intuitively, this describes the latent ability level at which a student would have equal chance of answering correctly and incorrectly. A benefit of the IRT framework is its representation of the interaction between

student ability level and item difficulty, so we also obtain estimates for each individual student's ability level. A histogram of student ability levels from the fitted model are plotted along the left-hand side, allowing the discriminability of the items to be compared to the predicted skill-levels of the students.

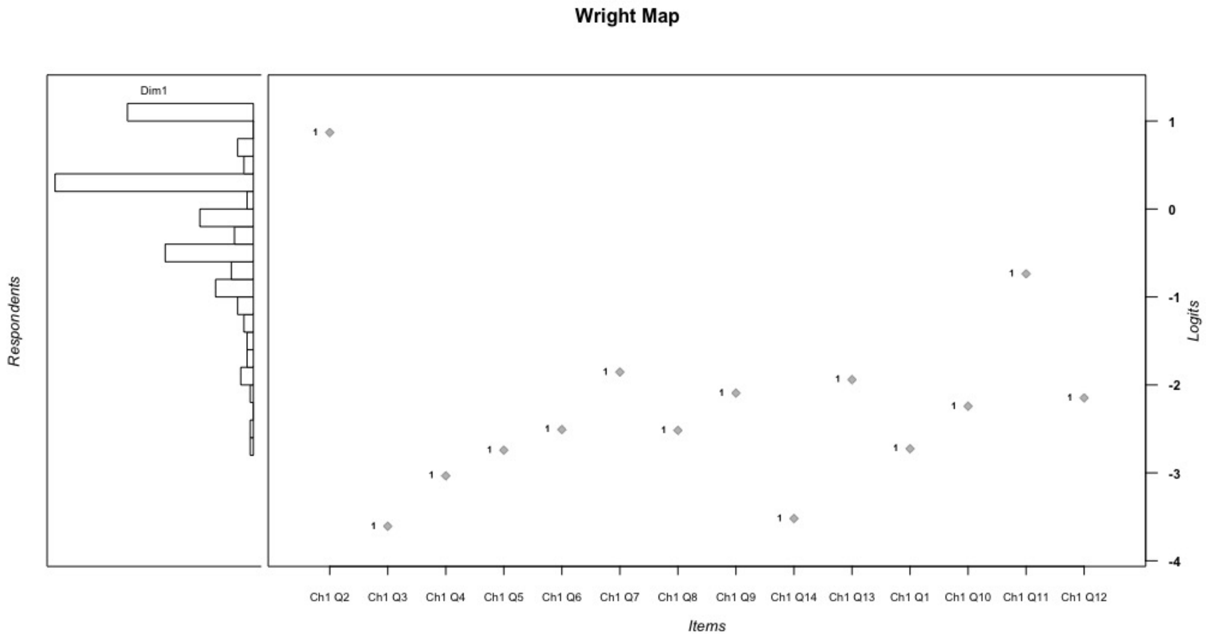

Figure 5.3: Wright Map for Chapter 1 review items

Overall, these results suggest a misalignment of the ability level of students and the difficulty level of the Chapter 1 review questions. Our model suggests that these questions are quite easy relative to the ability level of the students after reading and completing the formative activities in Chapter 1. It is important to note that this is not necessarily a bad thing in a traditional textbook. Such straightforward tasks reinforce the content of the text, clear up basic questions, and provide simple practice before applying that content in new settings.

However, the potential integration of LLMs to augment instruction and assessment forces us to reevaluate this type of rote practice in the light of the strengths, weaknesses, and concerns of AI tools discussed previously. A major benefit of AI-integration is the ability to tailor practice to individual needs, removing the top-down structure of hand-made problems. This is an element of the *CourseKata* textbook that is perhaps most suitable for automation. The evidence of AI-generated questions successfully capturing core material along with the trend of improved performance of students in computer-based learning environments on tasks requiring applications of concrete procedures together suggest that LLM tools could be applied well to personalize and improve engagement at these types of routine exercises. Moreover, this alleviates some of the difficult task of hand-tailoring a series of test items to measure across a wide range of ability levels and mathematical content, since students would be able to partially dictate which areas they receive additional practice.

However, there is likely still value in retaining hand-generated questions in some capacity. Content generated by LLMs or other automated question generation systems is highly likely to vary from student to student. So, groups of students who wish to work collaboratively are unlikely to see common questions. Since collaboration is well-documented to enhance mathematics learning [53], preserving some amount of hand-generated likely adds value. Moreover, question-generation resulting from interactions with LLMs carries no information about why those questions were produced in the first place due to the black box of transformer architectures, making both individual-level and class-level progress assessment difficult.

This suggests that there is value in augmenting LLM-based learning with hand-generated items. However, those items should be reworked to provide clearer indications of student progress. In the setting of our IRT model, this involves aiming for a greater variety of item difficulty parameters. Additionally, challenging questions have substantial pedagogical value in math learning [44], so there is likely value in introducing more difficult items even outside of the context of incorporating AI tools.

These hand-written items also allow for deeper additional analyses and insights into trends in student responses that would prove useful. Similar multidimensional analyses are likely worth performing on later chapters review instruments as well. In particular, clustering techniques that allow for detection of trends in student responses [20] may can help to indicate content areas where algorithmic question-generation and student-driven AI interaction will prove useful. Additionally, replacing or augmenting within-text items in *CourseKata* is also likely worth exploring. However, the tradeoffs discussed throughout this report should be taken into account both in selecting what tasks are well-suited for LLM interaction, but also how the user interface should be presented and perhaps evolved over the course of the textbook to suit individual tasks.

CHAPTER 6

Conclusion

The rapidly increasing capabilities of AI and NLP systems invites seemingly endless uses across a range of sectors. However, the application to teaching sits at a unique position. A truly automated teacher would require deep mathematical understanding, clear communication skills, social awareness, and situational consciousness to know when to simply watch and listen. Indeed, it is precisely the things that make teaching such a stimulating and rewarding enterprise for humans that makes it such a unique challenge for AI.

With the dizzying rate of performance improvement at tasks previously reserved for human reasoning capacities, traditional modes of instruction like university lecture and standard supplementary resources like textbooks seem bound for upheaval. However, as these AI systems inevitably work their way into classrooms, textbooks, and online tutorials, the inescapably social nature of teaching will also inevitably shape the way the relationship between human and machine cognition continues to evolve. Algorithmic instruction forces us to revisit the unique value that human instructors offer to their pupils in the first place. A hopeful future for AI in education is one in which these tools can be leveraged to help teachers curate more of the moments of tenderness, delight, and love that stay with their students for lifetimes.

REFERENCES

[1] Abdelghani, Rania, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 'GPT-3-Driven Pedagogical Agents for Training Children's Curious Question-Asking Skills'. *ArXiv* [Cs.CL], 2023. arXiv. http://arxiv.org/abs/2211.14228.

[2] Anderson, Jenny. "Push for A's at Private Schools Is Keeping Costly Tutors Busy". *The New York Times* (2011).

[3] Anderson, Richard C., and W. Barry, Biddle. "On asking people questions about what they are reading". *Psychology of Learning and Motivation* (1975): 89–132.

[4] Bashir, Amreen, Shahreen Bashir, Karan Rana, Peter Lambert, and Ann Vernallis. 'Post-COVID-19 Adaptations; the Shifts Towards Online Learning, Hybrid Course Delivery and the Implications for Biosciences Courses in the Higher Education Setting'. *Frontiers in Education* 6 (2021). https://doi.org/10.3389/feduc.2021.711619.

[5] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 'On the Dangers of Stochastic Parrots'. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 2021. https://doi.org/10.1145/3442188.3445922.

[6] Bender, Emily M., and Alexander Koller. 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data'. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.463.

[7] Bernstein-Yamashiro, Beth, and Gil G. Noam. 'Teacher-Student Relationships: A Growing Field of Study'. *New Directions for Youth Development* 2013, no. 137 (April 2013): 15–26. https://doi.org/10.1002/yd.20045.

[8] Boudreau, Emily. 'The Rapid Rise of Private Tutoring'. *Harvard Graduate School of Education*, May 2021. https://www.gse.harvard.edu/news/21/05/rapid-rise-private-tutoring.

[9] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 'Language Models Are Few-Shot Learners'. *ArXiv* [Cs.CL], 2020. arXiv. http://arxiv.org/abs/2005.14165.

[10] Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 'Sparks of Artificial General Intelligence: Early Experiments with GPT-4'. *ArXiv* [Cs.CL], 2023. arXiv. http://arxiv.org/abs/2303.12712.

[11] Chong Guan, , Jian Mou, and Zhiying Jiang. "Artificial intelligence innovation in education: A twenty-year data-driven historical analysis".*International Journal of Innovation Studies* 4, no.4 (2020): 134-147.

[12] Dodge, Jonathan, Q. Vera, Liao, Yunfeng, Zhang, Rachel K., Bellamy, and Casey, Dugan. "Explaining models".*Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[13] Drasgow, Fritz. "An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model".*Applied Psychological Measurement* 13, no.1 (1989): 77–90.

[14] Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. '"I Always Assumed That I Wasn't Really That Close to [Her]"'. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015. https://doi.org/10.1145/2702123.2702556.

[15] Farinde, Abiola A., and Chance W. Lewis. 'The Underrepresentation of African American Female Students in STEM Fields: Implications for Classroom Teachers', 2012.

[16] Fries, Laura, Ji Y., Son, Karen B., Givvin, and James W., Stigler. "Practicing connections: A framework to guide instructional design for developing understanding in complex domains".*Educational Psychology Review* 33, no.2 (2020): 739–762.

[17] Goldberg, Emma. "A.I.'s Threat to Jobs Prompts Question of Who Protects Workers".*The New York Times* (2023).

[18] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 'Generative Adversarial Nets'. In *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Vol. 27. Curran Associates, Inc., 2014.

[19] Guo, Wenjing, and Youn-Jeng, Choi. "Assessing dimensionality of IRT models using traditional and revised parallel analyses".Educational and Psychological Measurement 83, no.3 (2022): 609–629.

[20] Hu, Guanyu, Zhihua Ma, and Insu Paek. 'A Nonparametric Bayesian Item Response Modeling Approach for Clustering Items and Individuals Simultaneously'. *ArXiv* [Stat.AP], 2020. arXiv. http://arxiv.org/abs/2006.00105.

[21] Hu, Krystal. "CHATGPT sets record for fastest-growing user base - analyst note." (2023).

[22] Hwang, Gwo-Jen, and Ching-Yi Chang. 'A Review of Opportunities and Challenges of Chatbots in Education'. *Interactive Learning Environments* 0, no. 0 (2021): 1–14. https://doi.org/10.1080/10494820.2021.1952615.

[23] Jiang, Zhengbao, Jun, Araki, Haibo, Ding, and Graham, Neubig. ”How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering”. *Transactions of the Association for Computational Linguistics* 9 (2021): 962–977.

[24] Johnson, Carla C., Margaret J. Mohr-Schroeder, Tamara J. Moore, and Lyn D. English. *Handbook of Research on STEM Education*. Routledge, Taylor & Francis Group, 2020.

[25] Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. ‘ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education’. *Learning and Individual Differences* 103 (2023): 102274. https://doi.org/10.1016/j.lindif.2023.102274.

[26] Kelleher, Caitlin, Randy Pausch, and Sara Kiesler. ‘Storytelling Alice Motivates Middle School Girls to Learn Computer Programming’. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007. https://doi.org/10.1145/1240624.1240844.

[27] Kim, Tae Wan, and Bryan R. Routledge. ‘Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach’. *Business Ethics Quarterly* 32, no. 1 (2022): 75–102. https://doi.org/10.1017/beq.2021.3.

[28] Kizilcec, René F. ‘How Much Information?’ *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016. https://doi.org/10.1145/2858036.2858402.

[29] Koskinen, Rauno, and Harri Pitkäniemi. ”Meaningful Learning in Mathematics: A Research Synthesis of Teaching Approaches”. International Electronic Journal of Mathematics Education 2022 17 no. 2 (2022): em0679. https://doi.org/10.29333/iejme/11715

[30] Kosmyna, Nataliya, Cassandra, Scheirer, and Pattie, Maes. ”The Thinking Cap: Fostering Growth Mindset Of Children By Means Of Electroencephalography And Perceived Magic of Harry Potter Universe.” . *In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021.

[31] Kricorian, Katherine, Michelle Seu, Daniel Lopez, Elsie Ureta, and Ozlem Equils. ‘Factors Influencing Participation of Underrepresented Students in STEM Fields: Matched Mentors and Mindsets’. *International Journal of STEM Education* 7 (12 2020). https://doi.org/10.1186/s40594-020-00219-2.

[32] Krueger, Alan B. “Experimental Estimates of Education Production Functions.” *The Quarterly Journal of Economics* 114, no. 2 (1999): 497–532. http://www.jstor.org/stable/2587015.

[33] Krupa, Erin E., Corey Webel, and Jason McManus. ‘Undergraduate Students’ Knowledge of Algebra: Evaluating the Impact of Computer-Based

and Traditional Learning Environments'. *PRIMUS* 25, no. 1 (2015): 13–30. https://doi.org/10.1080/10511970.2014.897660.

[34] Lee, Min Kyung, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 'Procedural Justice in Algorithmic Fairness'. *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (2019): 1–26. https://doi.org/10.1145/3359284.

[35] Lee, Min Kyung. 'Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management'. *Big Data & Society* 5, no. 1 (March 2018): 205395171875668. https://doi.org/10.1177/2053951718756684.

[36] Lee, Min Kyung, and Katherine Rich. 'Who Is Included in Human Perceptions of Ai?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust'. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. https://doi.org/10.1145/3411764.3445570.

[37] Liangming Pan, , Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. "Recent Advances in Neural Question Generation." (2019).

[38] Lockhart, Paul. *A Mathematician's Lament*. Bellevue Literary Press, 2009.

[39] McLure, Felicity I., Barry J. Fraser, and Rekha B. Koul. 'Structural Relationships between Classroom Emotional Climate, Teacher–Student Interpersonal Relationships and Students' Attitudes to Stem'. *Social Psychology of Education* 25, no. 2–3 (2022): 625–48. https://doi.org/10.1007/s11218-022-09694-7.

[40] Mulwa, Ednah Chebet. 'Difficulties Encountered by Students in the Learning and Usage of Mathematical Terminology: A Critical Literature Review'. *Journal of Education and Practice* 6 (2015): 27–37.

[41] Neudert, Lisa-Maria, Aleksi, Knuutila, and Philip N., Howard. *Global Attitudes Towards AI, Machine Learning & Automated Decision Making.*Oxford Commission on AI & Good Governance, 2020.

[42] Niven, Timothy, and Hung-Yu Kao. 'Probing Neural Network Comprehension of Natural Language Arguments'. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–64. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1459.

[43] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 'Attention Is All You Need'. *ArXiv* [Cs.CL], 2017. arXiv. http://arxiv.org/abs/1706.03762.

[44] Papadopoulos, Ioannis. "Using tasks to bring challenge in Mathematics Classroom".Journal of Pedagogical Research 4, no.3 (2020): 375–386.

[45] Pataranutaporn, Pat, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. ‘Ai-Generated Characters for Supporting Personalized Learning and Well-Being’. *Nature Machine Intelligence* 3, no. 12 (2021): 1013–22. https://doi.org/10.1038/s42256-021-00417-9.

[46] Richardson, Christopher, and Larry Heck. ‘Commonsense Reasoning for Conversational AI: A Survey of the State of the Art’. *ArXiv* [Cs.CL], 2023. arXiv. http://arxiv.org/abs/2302.07926.

[47] Rouet, Jean-Francois, and Eduardo, Vidal-Abarca. ”“mining for meaning:” cognitive effects of inserted questions in learning from scientific text”.*The Psychology of Science Text Comprehension* (2014): 429–448.

[48] Sadik, Alaa. ‘Digital Storytelling: A Meaningful Technology-Integrated Approach for Engaged Student Learning’. *Educational Technology Research and Development* 56, no. 4 (2008): 487–506. https://doi.org/10.1007/s11423-008-9091-8.

[49] Silver, David, Aja, Huang, Chris J., Maddison, Arthur, Guez, Laurent, Sifre, George, Driessche, Julian, Schrittwieser, Ioannis, Antonoglou, Veda, Panneershelvam, Marc, Lanctot, and et al.. ”Mastering the game of go with deep neural networks and Tree Search”.*Nature* 529, no.7587 (2016): 484–489.

[50] Skinner, Ellen, and Michael Belmont. ‘Motivation in the Classroom: Reciprocal Effect of Teacher Behavior and Student Engagement across the School Year’. *Journal of Educational Psychology* 85 (12 1993): 571–81. https://doi.org/10.1037/0022-0663.85.4.571.

[51] Son, Ji Y., Adam B., Blake, Laura, Fries, and James W., Stigler. ”Modeling first: Applying learning science to the teaching of introductory statistics”.*Journal of Statistics and Data Science Education* 29, no.1 (2021): 4–21.

[52] Steuer, Tim, Anna Filighera, Tobias Meuser, and Christoph Rensing. ‘I Do Not Understand What I Cannot Define: Automatic Question Generation With Pedagogically-Driven Content Selection’. *ArXiv* [Cs.CL], 2021. arXiv. http://arxiv.org/abs/2110.04123.

[53] Sofroniou, Anastasia, and Konstantinos, Poutos. ”Investigating the effectiveness of group work in Mathematics”.Education Sciences 6, no.4 (2016): 30.

[54] Sultan, Md Arafat, Steven Bethard, and Tamara Sumner. ‘Towards Automatic Identification of Core Concepts in Educational Resources’. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 12 2014, 379–88. https://doi.org/10.1109/JCDL.2014.6970194.

[55] Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. ‘OLMpics – On What Language Model Pre-Training Captures’. *ArXiv* [Cs.CL], 2020. arXiv. http://arxiv.org/abs/1912.13283.

[56] Vallance, Chris. "AI could replace equivalent of 300 million jobs - report".*BBC* (2023).

[57] Vie, Jill-Jênn, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda. 'A Review of Recent Advances in Adaptive Assessment', 94:113–42, 02 2017. https://doi.org/10.1007/978-3-319-52977-6_4.

[58] Wang, Ming-Te, and Jessica L. Degol. 'Gender Gap in Science, Technology, Engineering, and Mathematics (Stem): Current Knowledge, Implications for Practice, Policy, and Future Directions'. *Educational Psychology Review* 29, no. 1 (2016): 119–40. https://doi.org/10.1007/s10648-015-9355-x.

[59] Wang, Tony T., Adam Gleave, Tom Tseng, Nora Belrose, Kellin Pelrine, Joseph Miller, Michael D. Dennis, et al. 'Adversarial Policies Beat Superhuman Go AIs'. *ArXiv* [Cs.LG], 2023. arXiv. http://arxiv.org/abs/2211.00241.

[60] Wu, Eric Hsiao-Kuang, Chun-Han Lin, Yu-Yen Ou, Chen-Zhong Liu, Wei-Kai Wang, and Chi-Yun Chao. 'Advantages and Constraints of a Hybrid Model K-12 E-Learning Assistant Chatbot'. *IEEE Access* 8 (2020): 77788–801. https://doi.org/10.1109/ACCESS.2020.2988252.