Artificial Test-Takers as Transformed Controls: Measuring SAT Difficulty Drift and Student Performance

Abstract

Introduction: Standardized test score trends are widely used to track student performance and inform policy, but they are difficult to interpret when exam content changes over time. We introduce an artificial test-taker framework that uses a fixed large language model as a stable benchmark to measure SAT Math difficulty drift and construct difficulty-adjusted measures of student performance.

Methods: We built a longitudinal SAT Math item bank from SATs spanning 2007–2023. For each year, we generated 50 bootstrapped SAT forms that match the year-specific section blueprint and administered all items to GPT-4 under a fixed set of parameters and training to develop counterfactuals. We combine our difficulty benchmarks with national SAT Math scores released by the College Board to assess robustness to compositional changes.

Results: The artificial test-taker framework indicates a statistically significant decline in SAT Math difficulty of 0.21σ relative to 2012. After adjusting for test difficulty using the transformed-control benchmark, student performance declines by 34 points in Average Difference in Scores (ADS) from 2012 to 2023. Heterogeneity analyses show that these declines are not uniform across racial groups.

Discussion: Artificial test-takers provide a scalable, protocol-invariant audit of longitudinal comparability when traditional equating is infeasible, opaque, or incomplete. Our findings imply that evolving SAT Math content can mask substantial underlying performance decline and can differentially obscure trends across student subgroups. More broadly, transformed-control designs using AI offer a tool for benchmarking educational outcomes and for separating changes in measured performance from changes in the measurement instrument itself.

Keywords: Artificial Test-Takers, Transformed Control, Large Language Models, Standardized Testing

1 Introduction

The Scholastic Aptitude Test (SAT) has long been central in U.S. college admissions, and researchers often use SAT scores to study educational outcomes, inequality and labor market outcomes (Card and Rothstein, 2007; Hoekstra, 2009). Despite the wide usage of standardized tests – such as the SAT – as a key metric for tracking student preparedness over time, interpreting trends naively can be misleading if changes in exam difficulty are not accounted for as SAT itself is not a static instrument – its content and difficulty can evolve over the years. When such comparability fails, raw score trends can systematically overstate or understate changes in underlying performance. Large-scale testing programs attempt to preserve comparability through linking and equating: statistical procedures that place scores from different forms onto a common scale so that they can be interpreted interchangeably (Kolen and Brennan, 2014; Holland and Dorans, 2006). Since forms differ in difficulty, the relationship between raw and scaled scores must be adjusted so that a given scaled score reflects the same performance level regardless of test date. This is the principle behind the College Board's public description of equating for the SAT, which explicitly frames equating as ensuring that "a score ... means the same regardless of when the student took the test" (College Board, 2019). In practice, however, equating and linking rest on strong requirements—most notably that forms measure the same construct and that the linking relationship is sufficiently invariant across populations and contexts—and these requirements can be stressed when tests and testing populations evolve (Michaelides, 2010; Guo et al., 2017).

A central threat to longitudinal comparability is *drift*: changes in items and construct representation over time. Even in well-run programs, items can become easier with exposure, change in ways that interact with curriculum and preparation, or shift in difficulty when the cognitive demands of the test evolve. Modern psychome-

tric work therefore treats drift as an empirical object that must be monitored, using statistical detection approaches designed for repeated or continuous testing environments (Guo et al., 2017; Lee and Lewis, 2021; Kang, 2023). Importantly, drift can matter even when operational equating is performed, because equating can only correct difficulty differences under the assumptions that justify the linking relationship and under the information available such as anchor items, recycled questions, and the representativeness of the linking design.

These issues are especially salient for the SAT over the past decade. The test has experienced consequential design changes, including the 2016 redesign[1] (College Board, 2016), while the surrounding admissions and participation environment has also shifted, particularly during the rapid expansion of test-optional policies in the COVID-19 era (Belasco et al., 2015; Bennett, 2021). At the same time, broader national indicators suggest that U.S. mathematics performance has declined in the pandemic period, including declines documented in NAEP long-term trend reporting (National Center for Education Statistics, 2023, 2024b,a). In this setting, interpreting SAT score trends requires separating at least two moving parts: (i) changes in the difficulty and construct representation of the SAT Math instrument, and (ii) changes in the performance and composition of the student population taking the test.

A practical obstacle is that researchers outside the testing agencies typically do not observe the operational equating and item-monitoring processes. Traditional equating designs rely on common anchor items or common examinees; these designs are difficult to implement for long-run audits when anchor items are proprietary and when it is challenging to administer historical forms to a stable sample of human examinees (Holland and Dorans, 2006). The common test-taker design is conceptually attractive because it reuses examinees rather than items, but it is rarely feasible at scale in high-stakes settings (Liu et al., 2025). As a result, applied research often treats SAT scores as directly comparable across years by assumption—an assumption that is convenient, but not always theoretically or empirically warranted when the instrument and participation context are changing.

In this paper, we introduce an artificial test-taker framework that uses a fixed

[1]Changes include no negative marking and restructuring of sections.

large language model (LLM) as a stable benchmark to quantify SAT Math difficulty drift and to construct difficulty-adjusted measures of student performance. Conceptually, our design can be understood as a *fixed artificial test-taker equating audit*: instead of holding items constant via anchors, we hold the examinee constant by repeatedly administering year-specific SAT Math content to the same artificial participant under invariant testing conditions. This produces a benchmark series capturing how difficult each year's form would be for a constant "test-taker", making it possible to distinguish changes in the instrument from changes in observed student scores. Methodologically, we then treat this benchmark as a transformed control, analogous to synthetic control approaches (Abadie and Gardeazabal, 2003), so that the difficulty-adjusted difference between student outcomes and the benchmark provides an interpretable measure of performance net of estimated difficulty drift (Callaway and Sant'Anna, 2021).

Our use of an LLM as the benchmark participant is motivated by two developments. The first is the rapid adoption of generative AI in assessment and educational measurement: recent literature argues that generative AI can serve as a diagnostic tool for item quality, validity evidence, and comparability, while also raising fairness and transparency concerns (Swiecki et al., 2022; Hao et al., 2024; Kaldaras et al., 2024; Bulut et al., 2024). The second is the emergence of methodological guidance for using LLMs as research objects or research participants, emphasizing the need for protocol specification, reproducibility, prompt transparency and careful reporting of model versioning and parameters (Chang et al., 2024; Abdelkarim et al., 2025; Zhao et al., 2025; Abdurahman et al., 2025). In line with these principles, we hold model version, prompting, and decoding fixed and treat each test-taker response as an independent interaction. We also conduct alignment and robustness analyses to assess whether model-perceived difficulty behaves in ways consistent with human difficulty signals and to evaluate concerns such as memorization.

Empirically, the artificial test-taker indicates a statistically significant decline in SAT Math difficulty of approximately 0.21 standard deviations relative to 2012. Accounting for this difficulty drift materially changes the interpretation of score trends. After difficulty adjustment, we estimate a larger decline in student per-

formance: about 34 SAT Math points from 2012 to 2023. We interpret these as difficulty-adjusted observational differences, as our design controls for estimated test difficulty drift but cannot fully eliminate confounding from shifts in participation or student composition, which are particularly plausible in the test-optional era (Belasco et al., 2015; Bennett, 2021). Our heterogeneity analysis reveals that these difficulty-adjusted declines in student performance are not uniform across racial groups.

By making the comparability problem explicit and by utilizing an artificial test-taker when traditional equating is infeasible or incomplete, we contribute to literature in both educational measurement and applied AI that uses standardised scores for inference. More broadly, transformed-control designs using AI offer a tool for benchmarking educational outcomes and for separating changes in measured performance from changes in the measurement instrument itself.

The remainder of the paper is structured as follows. Section 2 describes the SAT data used in our study and the GPT-4 test-taking process. Section 3 outlines our empirical strategy, including the transformed control method. Section 4 presents our main findings on SAT difficulty and student performance trends along with heterogeneity analysis. Finally, section 5 concludes with a discussion of the implications of our results for educational policy and future research.

2 Background & Data

2.1 SAT Questions

We constructed a question bank of SAT questions curated from a variety of online sources specializing in SAT preparation. These sources gave access to an extensive collection of past SATs. We gathered these tests in PDF format and stored them in a secure digital repository, encompassing a chronological collection from 2012 to 2023.

We transcribed each SAT exam PDF into a dataframe containing the question text, answer options for MCQs, source document, section and question numbers, question type, and calculator policy. MCQs present a set of possible answers, re-

quiring the test-taker to select the most appropriate option. Answer type allowed a one-line input from the test-taker[2].

Our final item bank contains 1,204 text-based questions from test forms spanning 2012–2023, representing 67% of all expected items. The remaining 33% of items contained visual elements incompatible with text-only models. Coverage is stable across years (range: 60–73%), suggesting no systematic temporal bias in which items could be transcribed. This question bank was then used to administer multiple SATs to the LLM by bootstrapping the examinations for each year. In order to stabilize annual estimates, we ran 50 independent bootstraps per year under an identical prompt and decoding protocol; all API calls were stateless and seed-controlled to ensure independence across trials.

In 2016, the SAT underwent significant changes. The format shifted from a total score of 2400 to 1600, aligning with the pre-2005 SAT format. The Math section retained its value of 800 points. Importantly, the revision eliminated the penalty for incorrect answers, encouraging students to attempt all questions without fear of point deduction for incorrect guesses. Additionally, the Math section saw a reduction in the number of sections from three to two and an increase in questions from 54 to 58. Before 2016, calculators were allowed for all math sections. Starting 2016, the Math section included both a calculator-permitted and a *no calculator* section. It is important to acknowledge these format changes and how they might impact our study. For further details, refer to the supplementary materials.

2.2 Student SAT Score Data

1. *National-level SAT scores data.* The College Board publishes yearly U.S. level reports summarizing SAT performance of the test-takers. The data from these reports include average scores, total test-takers and standard deviation in the scores for the mathematics test and the language and writing test. These reports are readily avail-

[2]Given high performing multi-modal AI models that accept visual prompts were not available at the time of running this experiment, questions incorporating graphical elements (figures, charts, diagrams) could not be transcribed for analysis. This exclusion ensured compatibility with the text-only input capabilities of the GPT-4 class models.

able year 2016 onwards, referred to as *post* period in our study, through the College Board website. For the years prior to 2016, which we refer to as the *pre* period, we collected the reports from the Internet Archive.

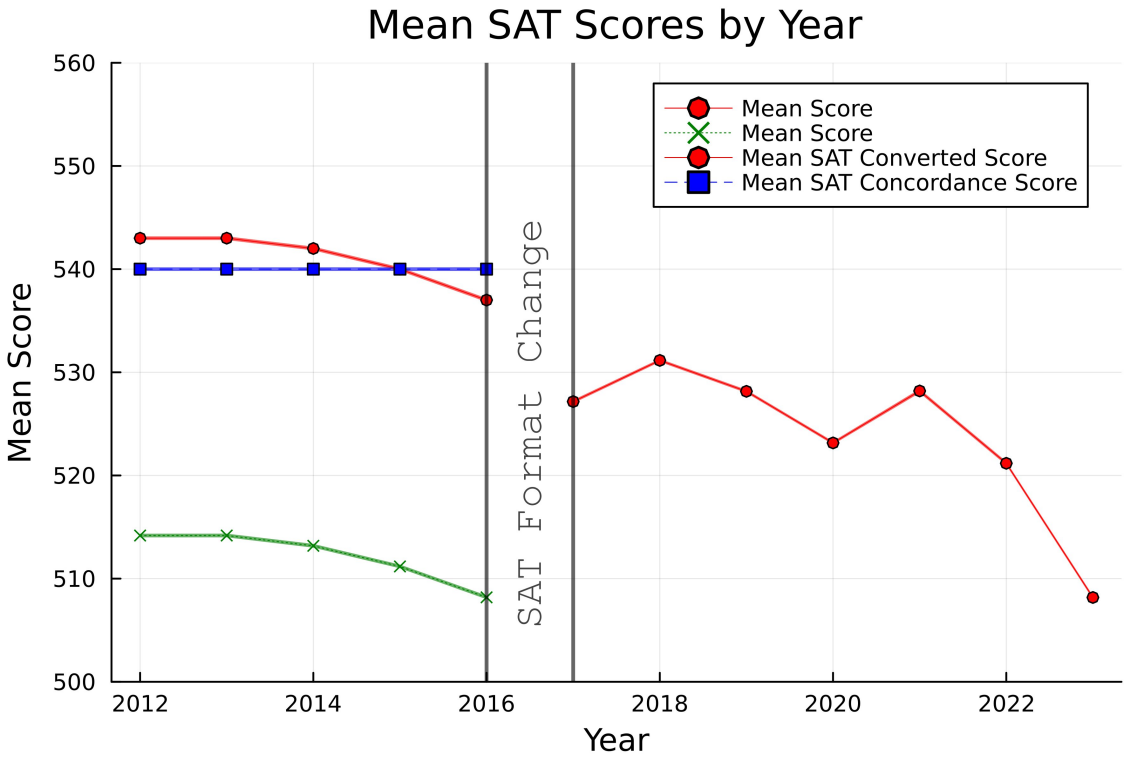


Figure 1: Average SAT Scores of test-takers

To account for the format change in the SAT exam in 2016, we used the concordance table provided by the College Board to convert the average SAT scores for the *pre* period to the *post* period. The concordance table provided by the College board is in multiples of 10. This requires rounding the average SAT score for exams in the *pre* period to the nearest multiple of 10 before it can be mapped to the average SAT score based on the concordance table. We used a linear model to interpolate the average SAT scores, which we call the SAT Converted scores. Figure 1 shows the average SAT score of students by year, before and after the conversion, as well as after using the linear interpolation. The green line represents the original average SAT

scores for each year from 2012 to 2016. The blue line represents the converted score obtained after rounding the scores in the green line to the nearest multiple of 10, and then applying the conversion in the concordance table. The red line before the SAT format change represents the converted score using the linear interpolation to match the scores after the format change. The average SAT scores in red for each year from 2017 to 2023 remain unaltered and are directly available through the College Board reports. After the conversion, we can see that the mean SAT score decline in the *post* period compared to the *pre* period. Figure 1 suggests that students are doing worse in the *post* period exams than the *pre* period exams, which also includes a significant Covid-19 effect. However, this is under the assumption that the underlying exam did not change in difficulty.

2. *State-level SAT scores data*. For state-level SAT reports from 2016 onwards, we directly accessed the state-level reports available on the College Board's website. These reports provide comprehensive insights into the SAT performance of students on a state-by-state basis, enabling a detailed examination of trends and patterns in SAT scores across the United States. However, the availability of state-level SAT data prior to 2016 posed a unique challenge. To procure this historical data for the *pre* period (2012-2016), we used the Internet Archive. All human data are aggregate; no identifiable records were accessed; generative AI was used as an experimental agent, not as an author (separate disclosure provided).

3. *District-level SAT scores data*. We used publicly available school district level SAT data for the state of Massachusetts from the Massachusetts Department of Education. The data are available for all the school districts in the state for each year from 2004. The number of test-takers, the average reading, writing and mathematics scores are provided by school district for a given academic year. We identified 228 unique school districts commonly present over all the years considered in the study. For further details, refer to the supplementary materials. We present two national views—(i) population-weighted national reports and (ii) a state unweighted average—and additionally a district-level panel (Massachusetts) to check robustness to composition.

2.3 Artificial participant

We used GPT-4-0125-preview as the artificial participant with fixed decoding (temperature = 0, max_tokens = 100); calls were executed in January–February 2024. Each year's evaluation comprised 50 bootstraps, each sampling a balanced SAT exam from our question bank from each year; outputs captured include final answer, correctness. Chain-of-thought was disabled, and answers were formatted deterministically for machine scoring. No further use of AI was made beyond this evaluation in the analysis.

3 Empirical Strategy

In order to estimate the change in difficulty of the SAT math section relative to the baseline period, we extend the principles of Synthetic Control by Abadie and Gardeazabal (2003). LLMs like OpenAI's ChatGPT have revolutionized Natural Language Processing (NLP) tasks. Here, the transformed control is built by holding the test-taker fixed (GPT-4 under an identical protocol) and comparing year-to-year performance on contemporaneous items, so the counterfactual uses each year's own SAT content and precludes extrapolation.

We do not explicate the LLMs further, Korinek (2023) and Dell (2024) provide a comprehensive guide on their day-to-day use case. As these models are trained on increasing amounts of human generated text data and are continually improved, they become more adept at understanding and responding to queries. This progress opens new and sophisticated applications for these models. We highlight one such application within social sciences: what we refer to as transformed control in our research design. The LLMs have been used for direct causal reasoning by Kiciman et al. (2023), given a causal question to understand the models' capacity to accurately identify causal factors. Our approach is distinct from this procedure as we are using the LLM to generate control group data to enrich the inference about the student test-takers' performance. Although the control in this design does not serve for causal inference yet, it paves the way for incorporating modern AI tools into social

science research.

The transformed control has similar advantages that are observed with synthetic control methods highlighted by Cunningham (2021). Our method precludes extrapolation, the comparison with student test-takers' SAT outcomes is based on the LLM's performance in SAT questions from the same year. The counterfactual constructed relies on questions students themselves faced in that year. The LLM does not learn from previous attempts as each API call to the model is independent. This ensures the LLM's responses are not influenced by peeking at the student outcomes as each evaluation of the SAT is independent. Instead, it uses the knowledge in its training data to output an appropriate answer to the question provided. While standardized tests have been used to identify trends in student performance over time, it is assumed the underlying exam is uniformly challenging. Independently evaluating the standardized tests would require significant logistical undertaking. The LLM can instead act as the control and thus bridge the gap between qualitative and quantitative research. In our experiment the LLM by OpenAI, GPT-4 is the control unit.

3.1 Scoring methodology

The section outlines the methodology employed to enable OpenAI's models to answer SAT questions. As illustrated in Figure 2a, we randomly sampled 50 SATs for each year from the question bank using bootstrapping while maintaining identical number of question types as per the SAT format of the given year. We then prompted GPT-4 by OpenAI to take each of the 50 sampled exams for each year. Then, we calculated the number of correct answers and their proportion (number of correct answers divided by total questions) to assess the model's performance and compared the distributions obtained for different periods. We keep model version and decoding fixed across all calls and treat each API call as stateless and independent, ensuring protocol invariance across years.

First, we sample an SAT without replacement from our SAT Math Question bank as in Section 2. Next, we proceed by selecting individual questions from the sampled SAT exams, each of which is then presented to the GPT-4 model using a

carefully pre-formulated prompt. The structure of our prompt has been provided in the supplementary materials. Once GPT-4 receives this API call, it provides a response in a format specified by our prompt. The answer is then saved to a dataset that keeps a record of all the questions that have been answered and their corresponding answers. This process continues until all the questions in the sampled exam have been exhausted. All the questions and answers associated with each exam are compiled into a dataset. Once all the generated answers are collected, we convert raw score to scaled SAT score and recenter for accurate comparison over time, as described in Figure 2b and explained in Section 2.2.

The College Board accounts for variation in exam difficulty before scoring the students' performance within that year. As some students are likely to receive relatively more difficult exams than their peers taking the exam in the same year, the final score factors in the difficulty level to allow the accurate assessment of students relative to students in the same peer group. The process of bootstrapping as described generates exams within the year that include questions of varying difficulty, mimicking the exams received by student test-takers. Thus, some sampled exams are more difficult than others. Thereby, ensuring we evaluate the average SAT for a given period.

3.1.1 Prompting

We employed a zero-shot prompt, meaning no example solutions were provided to the LLM to facilitate answering the question. Further, we did not allow for chain-of-thought. The LLM was proscribed from sequential reasoning. We use the same model build [gpt-4-0125-preview] and unchanged prompting/decoding parameters for all years evaluated. The question was provided through the prompt and the LLM was asked to provide one character letter output for multiple choice questions and the appropriate numerical or equation output for the answer type questions. The prompting strategy remained identical over all periods, ensuring no bias through the prompt. Since all prompts are independent API calls to the model, the LLM had no memory of previous questions. The LLM therefore is tasked to answer each

question independent of any other question or reference to the period from which the question is gathered. This ensured the comprehensive evaluation of the difficulty of the questions from appropriate periods. Structured outputs to refine the LLM response were not available at the time of the study. Further prompting details and GPT-4 model parameters are provided in the supplementary materials.

3.1.2 Question Embeddings and Difficulty Alignment

To ensure that the composition of SAT questions in our study accurately reflects the intended distribution of difficulty levels, we perform a difficulty alignment analysis to ensure no over- or under-representation of any particular difficulty type. Further, having identified the difficulty of the questions across different years, we can assess the performance of the LLM on questions of varying difficulty. Below, we describe the methodology used and the results of this analysis [3].

For the exam years before 2014, the College Board provided difficulty ratings for each question. These difficulty ratings were not provided by the College Board for exams starting 2014, hence we predicted the difficulty rating using a machine learning model. To avoid temporal leakage, training uses only preserved 2008–2011 items and pre-2014 labels; 2015+ items are prediction-only and never used for training. The difficulty rating of the question for the students was between 1 and 5, with 1 being the easiest and 5 being the hardest. We used these ratings to study whether LLM's concept of difficulty aligns with the student test-takers. For this task, we used the text-embedding-3-large model from OpenAI (Neelakantan et al., 2022). Embeddings are a way to represent text as a vector in high dimension, which contains the semantic meaning of the text. These embeddings show how the LLM identifies the text in its internal representation and has been used for various language processing tasks (Xue et al., 2018; Gao et al., 2019).

The embeddings created by the model are used as features to predict the difficulty of the questions. In addition, the SAT was a pencil and paper test for the years considered in this study, and the difficult questions appear at the end of each

[3]We were able to use preserved SATs from 2008 to 2011, which were not a part of our final results, but serve as training data for the machine learning model in this analysis.

section. We use this information as an additional feature in the classification model by including a progress bar as a numerical representation of the progress through the section. To simplify the classification task, we group difficulty ratings 1 and 2 as easy (310 questions), 3 as medium (265 questions), and 4 and 5 as hard (205 questions). The model is trained on 80% of the data having balanced the classes and tested on the remaining 20% of data. We also balance classes and report out-of-sample performance on the held-out 20% split to verify alignment. Using a Random Forest Classifier, the model achieves an accuracy of 0.789 on the testing data. Additional details regarding the confusion matrix for the classification are available in supplementary materials. The model is then used to predict the difficulty of the questions in the SAT from 2015 to 2023.

Figure 3(A) displays the results, and highlights the proportion of easy, medium, and hard questions in the bootstrapped SATs from each year. Further, the average proportion of easy, medium, and hard questions in the complete SATs gathered for the years 2008 to 2014 were 40%, 34% and 26% respectively. The proportion of easy, medium, and hard questions in the bootstrapped SAT approximates the average proportion in the complete SATs, suggesting that years 2015 onwards had similar proportion of easy, medium, and hard questions when juxtaposed against the preceding years. These results suggest that SATs from varying years had similar proportions of easy, medium, and hard questions. This is important as it allows us to compare the performance of the LLM across years without worrying about the composition of the SATs.

Having predicted the difficulty of the questions, we use the predicted classes to identify the change in performance of the LLM, i.e., the proportion of each question type LLM is correctly answering over time. In Figure 3(B), we show the correct response rate of the LLM for easy, medium, and hard questions. First, we notice the LLM performs better on easy and medium difficulty questions compared to hard questions for the years 2008 to 2014. The LLM's lower accuracy rate on questions rated difficult for student test-takers suggests a potential alignment in perceived difficulty between students and the LLM. Additionally, while performance improves across all difficulty levels over time, the most pronounced gains are observed for hard

questions.

3.2 Reduced Form

For this study, we have two parameters of interest. First, we want to estimate the change in difficulty of SAT math section. Second, after controlling for exam difficulty, we want to estimate the change in performance of test-takers. We explain our strategy to estimate each of these parameters below. We fix 2012 as baseline and report changes relative to that year for both the model and students, so that the Average Difference in Scores (ADS) compares like-for-like movements from a common starting point.

1. *Transformed Control.* To estimate the change in difficulty of the SAT, we formally consider the question bank Q of 1,204 SAT questions from the years 2012 to 2023. S_t is a collection of subsets of Q and each S_t contains randomly sampled exams from the question bank Q without replacement, ensuring unique questions populate each exam. Each exam $E_{j,t} \in S_t$ and each $E_{j,t} \subset Q$, where $j = 1, 2, ..., 50$. $\forall E_{jt} \in S_t$, with $t = \{2012, 2013, ..., 2023\}$ and where q_{ijt} is a question in exam E_{jt}, we get

$$n_{mjt} = \sum_{i=1}^{|E_{jt}|} \chi_{A_{q_{ijt}}}(f_m(q_{ijt})) \text{ and } p_{mtj} = \frac{n_{mjt}}{|E_{jt}|},$$

n_{mjt} is the number of questions answered correctly by model m, for exam j in period t. p_{mjt} is the proportion of questions answered correctly by model $m \in \{\text{GPT-4 Turbo, Claude 3.5}\}$, for exam j in period t. While we restrict this study to GPT-4 model, we include results from other models in the supplementary materials. $|E_{jt}|$ is the cardinality of E_{jt}, which in our paper represents the number of questions in exam E_{jt}. f_m is some LLM based function that takes in a question as input and provides an answer as output, depending on whether the answer is for MCQ or answer-type question. $\chi_{A_{q_{ijt}}}$ is a characteristic function that is 1 if output from f_m is correct and 0 otherwise, depending on a set of correct answers $A_{q_{ijt}}$ for question q_{ijt}.

From these performance evaluation measures n_{mjt} and p_{mjt}, we obtain the statistics μ_{mt} and ν_{mt}, along with their respective standard errors $\sigma_{\mu mt}$ and $\sigma_{\nu mt}$ -

$$\mu_{m,t} = \frac{1}{|S_t|} \sum_{j=1}^{|S_t|} n_{mjt}$$

$$\nu_{m,t} = \frac{1}{|S_t|} \sum_{j=1}^{|S_t|} p_{mjt}$$

μ_{mt} is the mean number of questions answered correctly by model m in period t. ν_{mt} is the mean proportion of questions answered correctly by model m in period t. $|S_t|$ is the cardinality of the set S_t, which is 50 for all time periods.

2. *Performance Comparison.* Our second parameter of interest is the change in mathematical performance of the test-takers. To estimate this parameter, we utilize the estimation procedure from multi-period difference-in-difference framework as expounded in Callaway and Sant'Anna (2021). Next, we explain how we borrow ideas and notation from Callaway and Sant'Anna (2021) to fit our empirical setting.

The LLM, GPT-4 is a fixed entity given the training and prompting technique are held constant. If the test-takers did not change in composition, the expected change in performance of the students in the SAT exam is the estimated change in the underlying exam difficulty. We use the potential outcome notation for the test-takers' SAT score measurements, $Y_{i,t}(0)$ is potential i^{th} measurement in period t for the unchanging test-takers. While $Y_{i,t}(g)$ is the actual outcome measured for students. We formally define the j^{th} exam score for the LLM taken in period t as $T_{j,t}$. This allows us to construct a strong parallel trends assumption as the following,

$$\mathbb{E}[Y_t(0) - Y_{t-1}(0)] = \mathbb{E}[T_t - T_{t-1}], \text{ where } 2012 < t \leq 2023.$$

Since our transformed control is the unobserved parallel trend, the estimation mimics treatment effect estimation in a classical difference-in-difference approach without making causal claims. Additionally, under this parallel trends like assumption, the expected change in the LLM's score is the estimated change in difficulty of

the SAT exam between the measurement periods. We borrow some notation from the estimation procedure described by Callaway and Sant'Anna (2021) and adopt for our use case. Our control provides a richer perspective about the difference in student performance, we refer to this estimand as the Average Difference in Scores (ADS).

$$ADS(\hat{t}, t) = \mathbb{E}[Y_t(g) - Y_t(0) \mid \hat{t}\ ]$$

In the above equation, $Y_t(0)$ is the unobserved potential outcome. However, when we impose strong parallel trends, we can identify the ADS with reference to a baseline year,

$$\begin{aligned} ADS(\hat{t}, t) &= \mathbb{E}[Y_t(g) - Y_t(0) \mid \hat{t}\ ] + \mathbb{E}(Y_{\hat{t}}(0)|\hat{t}) - \mathbb{E}(Y_{\hat{t}}(0)|\hat{t}) \\ &= \mathbb{E}[Y_t(g) - Y_{\hat{t}}(0)|\hat{t}] - (\mathbb{E}[Y_t(0) - Y_{\hat{t}}(0)|\hat{t}]) \end{aligned}$$

Now, notice that because we start at the same baseline, we have $Y_{\hat{t}}(0) = Y_{\hat{t}}(g)$. So,

$$\begin{aligned} ADS(\hat{t}, t) &= \mathbb{E}[Y_t(g) - Y_{\hat{t}}(0) \mid \hat{t}\ ] - \mathbb{E}[Y_t(0) - Y_{\hat{t}}(0) \mid \hat{t}\ ] \\ &= \mathbb{E}[Y_t(g) - Y_{\hat{t}}(g) \mid \hat{t}\ ] - \mathbb{E}[Y_t(0) - Y_{\hat{t}}(0) \mid \hat{t}\ ] \\ &= \mathbb{E}[Y_t(g) - Y_{\hat{t}}(g) \mid \hat{t}\ ] - \mathbb{E}[T_t - T_{\hat{t}} \mid \hat{t}\ ] \end{aligned}$$

The above expression is identifiable as all the terms are observed. Hence, using strong parallel trends like assumption and fixing the baseline at some $\hat{t}$, we can estimate ADS.

Our analysis is operationalized through a standard parametric linear regression model that accommodates multi-valued discrete treatment variable, which can be represented by the following regression equation:

$$\Delta Z_{i,t,s} = \sum_{t=2013}^{2023} \mathbf{1}\{\tau_i = t\}\gamma_t + \sum_{t=2013}^{2023} \mathbf{1}\{\tau_i = t\} \times \mathbf{1}\{s = student\}\beta_t + \epsilon_{i,t,s}$$

In our study, $\Delta Z_{i,t,s}$ is the change in SAT score for unit i from the baseline year 2012, where $s \in \{student, LLM\}$ and τ_i is the exam year of unit i. The parameter β_t

is the $ADS(\hat{t}, t)$ measuring expected change in test-takers' math performance relative to baseline 2012 controlling for the exam difficulty through the LLM.

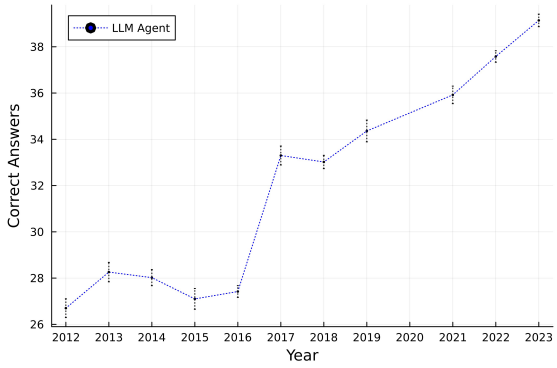
4 Results

We first present the performance of GPT-4 in the SATs over the years. Next, we use GPT-4's performance as a control to estimate the changes in student performance.

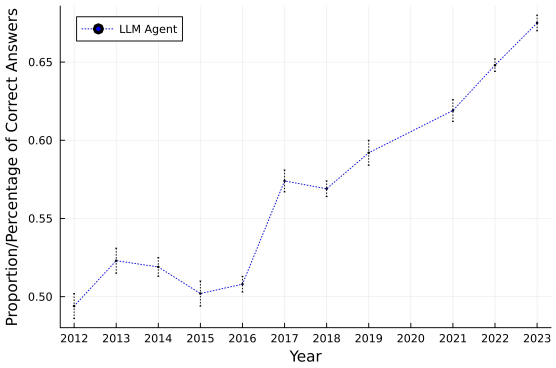
4.1 Change in Difficulty of SAT Math Section

In evaluating GPT-4's performance on SAT math sections, we accounted for the varying number of questions in the *pre* format change (54 questions) and *post* format change (58 questions) periods. To enable a fair comparison, we also analyzed the proportion of correctly answered questions for each year, i.e., the ratio of number of correct answers to total questions per exam over 50 bootstrap samples. This approach ensures a normalized comparison despite the different total question counts. This normalization lets us interpret year-to-year differences as changes in form difficulty rather than artifacts of question counts.

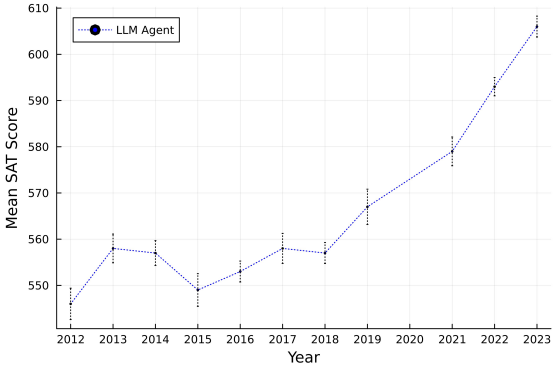
4.1.1 Average Performance of LLM on SATs



(A) Number of questions correctly answered by LLM in SATs from each year



(B) % of questions correctly answered by LLM in SATs from each year



(C) Scaled SAT scores for LLM in SATs from each year

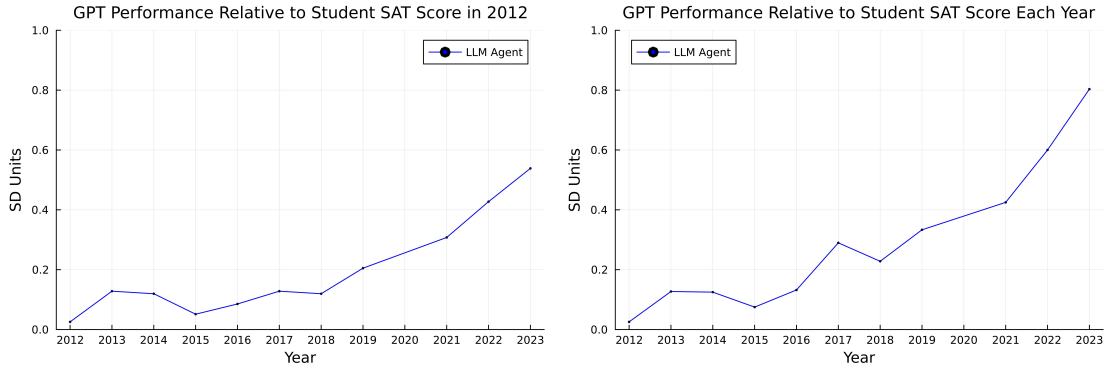
Figure 4: Performance of LLM on SAT exams over the years. Each point represents the mean across 50 bootstrap samples per year. Shaded regions show 95% bootstrap confidence intervals (2.5th–97.5th percentiles). Panel (A) shows raw correct answers, (B) shows proportion correct, and (C) shows scaled SAT scores (converted to post-2016 scale for years prior to 2017).

Figure 4(A) shows the raw number of questions correctly answered by the LLM in the SATs from each year. The LLM's performance is found to be increasing over time. As there are 4 additional questions after the format change, it is challenging

to interpret the raw number of questions correctly answered. Thus, Figure 4(B) provides the corresponding percentage of questions correctly answered by the LLM in the SATs from each year. Figure 4(C) provides the scaled SAT score which can be benchmarked against maximum SAT score of 800. We use the concordance tables provided by the College Board to recenter the scores before the format change as described in Section 2. As shown, GPT-4's performance is increasing over time. Since 2012, the SAT has been declining in difficulty by roughly 4 points per year.

4.1.2 Dispersion over the years

We also analyze the dispersion of GPT-4's performance relative to the average student performance in 2012. This analysis provides insights into how the LLM's performance has evolved over time compared to the average student test-taker. Specifically, we compared the performance of the LLM on the SAT to the national student score distribution from 2012, using this fixed year as a baseline to assess the LLM's relative standing. Additionally, we performed a year-by-year comparison, evaluating the LLM's performance relative to the national student score distribution for each corresponding year. In both analyses, we quantified the gap between the LLM and the average student in terms of the standard deviation of student scores, providing a standardized measure of the difference in performance over time.



(A) GPT performance relative to student score in 2012

(B) GPT performance relative to Student Score for each year

Figure 5: GPT performance relative to Student Scores

Figure 5(A) shows the performance of the LLM relative to the student score distribution in 2012. As shown by this figure, the LLM's average performance is deviating further away from the average student performance over the years, and moving closer to the right tail of the student score distribution[4]. In 2012, the gap between the LLM and the average student was 0.02σ. However, by 2023, this gap increased to 0.6σ, and the average over 12 years is 0.21σ relative to the base year of 2012. These standardized gaps provide effect-size comparability and are computed using the 2012 student SD (117) as noted.

Figure 5(B) shows the performance of the LLM relative to the student score distribution for each year[5]. The gap between the LLM and the average student is also increasing over time and even more pronounced, indicating that the LLM's performance is diverging from that of the average student. In 2012, the gap was 0.02σ, and by 2023, it had increased to 0.8σ, with an average of 0.32σ over the 12 years.

[4]Mean student score in 2012 was 543 after concordance and standard deviation was 117. We used the following formula: $\frac{\mu_{\text{Student},2012} - \mu_{\text{LLM,t}}}{\sigma_{\text{Student},2012}}$

[5]We used the following formula: $\frac{\mu_{\text{Student,t}} - \mu_{\text{LLM,t}}}{\sigma_{\text{Student,t}}}$

4.2 Change in Student Performance

We begin by providing results for estimated $ADS(\hat{t}, t)$ for the baseline year 2012. This gives us the relative change in the performance of average student test-taker, having controlled for the SAT difficulty using transformed control. We perform this comparison at the national level using both the population performance data provided by the College Board and the state-by-state unweighted average SAT scores to replicate a representative national sample. For the state of Massachusetts, we use the school district-level SAT scores to estimate ADS.

Figure 6(A) shows the estimated ADS, which gives us decline in SAT scores of student test-takers after controlling for exam difficulty. From this figure, we notice that the average performance of the test takers declined by 34 points from 2012 to 2023. Table 1 summarizes the main results: GPT-4's score increased by 59.8 points (SE = 4.68) from 2012 to 2023, while student scores declined by 35 points (SE = 0.13), yielding an ADS of 94.8 points (SE = 4.68). This indicates that the difficulty-adjusted decline in student performance is substantially larger than the raw score decline suggests.

Table 1: Main Results: GPT-4 and Student SAT Math Performance (2012–2023)

Year	GPT-4 Score	(SE)	Student Score	(SE)	ADS	(SE)
2012	546.2	(3.40)	543	(0.09)	0.0	—
2016	553.0	(2.27)	537	(0.10)	12.8	(4.09)
2020	597.2	(3.16)	523	(0.08)	71.0	(4.64)
2023	606.0	(3.21)	508	(0.09)	94.8	(4.68)
Change (2012 → 2023)						
	+59.8	(4.68)	−35	(0.13)	+94.8	(4.68)

Notes: GPT-4 scores from 50 bootstrap samples per year (SE = SD/$\sqrt{50}$). Student scores are national population means converted to post-2016 scale. ADS = (GPT-4 change) − (Student change), relative to 2012 baseline.

We used the state-by-state unweighted average SAT scores to construct a representative national sample along with the national population data provided by the

College Board. The ADS using both samples are shown in Figure 6(B) along with the district-level scores for the state of Massachusetts. Figure 6(C) shows the estimated ADS for Asian, Black and White students. White students show the largest decline in average performance of 33 SAT points, while Black and Asian students declined by 25 and 15 points. The College Board reports do not have consistent demographic information over the period considered for Hispanic students, we therefore considered the most consistent demographic breakdown of the SAT scores present in the report.

These results highlight key trends; state-by-state unweighted SAT data generally mirrors national population trend. Massachusetts presents a slightly different narrative, with changes in SAT scores that often align with national trends but with some exceptions. For instance, the decline in Massachusetts is less pronounced in certain years, indicating possible state-specific factors that might have cushioned the impact. Overall, we find a notable difference in performance of the LLM and the performance of test takers, with LLM's performance improving and student performance declining over time. These estimates are qualitatively invariant between the national population, the state unweighted average, with Massachusetts serving as a district-panel check that largely follows the national trend. Taken together, a steady easing in form difficulty coupled with a difficulty-adjusted decline in student performance underscores the value of an artificial test-takers' benchmarking when comparing cohorts over time.

4.3 Robustness Checks

We conduct several robustness checks to address potential concerns about our methodology.

Memorization Test. A key concern is whether GPT-4's improving performance reflects memorization of older SAT items that may appear in its training data (cut-off: September 2021). To address this, we administered paraphrased versions of 203 questions from years 2011, 2015, 2019, and 2023 to the same model. If memorization drove performance, paraphrasing should substantially reduce accuracy. Instead, as shown in Figure 7(A), we find only a 3.9 percentage point drop in accuracy (64.0%

to 60.1%), and crucially, the correlation between original and paraphrased accuracy across years is $r = 0.991$. This near-perfect preservation of the temporal trend strongly suggests GPT-4 is solving problems rather than retrieving memorized answers.

2016 Format Change. The SAT underwent a major redesign in 2016. To test whether this format change confounds our difficulty estimates, we conducted an event study examining whether a discrete jump occurs at the 2016-2017 boundary. As shown in Figure 7(B), the jump coefficient is statistically insignificant ($\beta = -9.82$, SE $= 11.96$, $p = 0.433$), indicating no discontinuity at the format change. The difficulty trend is smooth across the redesign, suggesting our results are not artifacts of format differences.

Question Type Consistency. We examine whether the difficulty trend is driven by particular question types. Both multiple choice questions (MCQ) and student-produced response (grid-in) questions show nearly identical improvement: MCQ accuracy increased by 21.4 percentage points (43.6% to 65.0%) and grid-in accuracy increased by 21.4 percentage points (37.8% to 59.2%) from 2008 to 2023. This consistency across question formats supports the robustness of our difficulty estimates.

Calculator vs No Calculator. The post-2016 SAT introduced a distinction between calculator-permitted and no-calculator sections. We assess the possibility that calculator policies might drive difficulty trends and find an insignificant performance difference: GPT-4 achieves a mean accuracy of 59.5% (SD = 4.8%) on Calculator sections versus 56.9% (SD = 5.8%) on No Calculator sections. The relatively small gap ($\sim$2.6 percentage points) in accuracy suggests that calculator policies are not a primary driver of the observed difficulty decline.

Difficulty Composition. Using a Random Forest classifier trained on pre-2014 College Board difficulty labels, we examine whether the composition of easy, medium, and hard questions changed over time. As shown in Figure 3, we find modest compositional drift: the proportion of hard questions shows a positive trend ($r = 0.62$), while easy questions show a negative trend ($r = -0.44$). However, the year-to-year variance is small (0.001–0.002), and there is no abrupt change at 2016, suggesting difficulty composition changes are gradual rather than discrete.

5 Discussion and Conclusion

Standardized tests are routinely used to track performance trends, study inequality and downstream outcomes, but such comparability is reasonable only when the measurement instrument is stable or when test forms are credibly linked onto a common scale. In psychometric terms, linking and equating are intended to preserve score comparisons across administrations by adjusting for form difficulty, such that the same reported score corresponds to the same performance level (Kolen and Brennan, 2014; Holland and Dorans, 2006). However, when the conditions that justify these adjustments are stressed—because the assessment evolves, populations shift, or anchor information is unavailable—raw score trends can misstate changes in underlying performance (Michaelides, 2010). We study this comparability problem directly by introducing an "artificial test-taker" benchmark that makes difficulty drift measurable, and by using our benchmark as a transformed control to help separate changes in measured performance from changes in the measurement instrument.

Our primary result suggests that SAT Math difficulty is not constant over time, as measured by a fixed artificial participant exposed to year-specific items under invariant testing conditions. Two features of this benchmarking result are particularly consequential for interpretation. First, the benchmark changes are systematic rather than idiosyncratic, suggesting a time pattern consistent with drift rather than noise. Second, the benchmark is constructed by holding the test-taking protocol fixed, which clarifies that the estimated drift is driven by differences in test content rather than changes in the agent's "testing environment". These findings align with the broader measurement literature that treats item drift as an empirical object to be monitored, not an assumption to be taken for granted (Guo et al., 2017; Lee and Lewis, 2021; Kang, 2023).

Our second key result uses the benchmark series to re-express observed student score trends after adjusting for test-difficulty. We find that the difficulty-adjusted decline in student performance is larger than as suggested by the raw score decline, implying that raw score trends that do not account for evolving exam difficulty may understate the decline in student performance. This mismeasurement is likely

complicated by concurrent shifts in student participation due to widespread test-optional policies (Belasco et al., 2015; Bennett, 2021). More broadly, recent work on assessment in the age of generative AI emphasises that the deployment of AI tools can both reveal and obscure inequities, depending on transparency, validation, and the interpretability of outputs (Swiecki et al., 2022; Kaldaras et al., 2024; Hao et al., 2024; Bulut et al., 2024). Finally, our results also indicate that difficulty-adjusted trends are not uniform across racial groups. The heterogeneity analysis underscores that comparability is not only a psychometric concern but can also affect how trends are perceived across subpopulations. In a setting where participation and incentives may shift differentially across groups over time, group-specific trajectories should be interpreted with care due to potential selection effects and changing composition of test-takers (Card and Rothstein, 2007).

Although in this study we perform supporting analysis to show question difficulty alignment between the LLM and the student test taker, the concept of alignment is still up for exploration and debate among Machine Learning and AI researchers (Bai et al., 2022; Ji et al., 2025). The LLM's lower accuracy on questions that student test-takers find difficult suggests that both the LLM and the test-takers perceive and are challenged by similar aspects of the questions. There likely exist factors that are influencing the LLM's performance that also affect human difficulty perception, such as question complexity and required reasoning skill. These factors can be further explored to understand the alignment between the LLM and the test-takers, but is beyond the scope of this study and may be an avenue for future research.

The LLMs in this study were not tuned to operate as a typical high school test taker; imbuing the LLMs with such a personality could offer stronger alignment. Further psychometric analysis of the LLM's performance on the SAT may provide insights into the detailed performance of the LLM. Multimodal AI models with visual inputs could analyze the SAT math section more thoroughly, these models were not available at the time of this study. These models can analyze image input but can misidentify the numerical data present in a graph or a chart, which has to be prompted separately adding to the complexity. While the National Assessment of Educational Progress (NAEP) has also noted a decline in math performance in recent

years (National Center for Education Statistics, 2023) and a widening academic achievement gap between the rich and the poor (Reardon, 2018), the NAEP questions and the American College Testing (ACT) questions were not available to perform a similar comparative study. Despite these hindrances, our study provides a template for the use of AI as an artificial test-taker to develop counterfactuals for evaluating long-term trends in educational outcomes. As agentic and fine-tuned AI models continue to improve, they will enable increasingly better approximations of human test-taking behavior, ultimately yielding more robust counterfactuals for longitudinal educational research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Replication code for the results can be found on GitHub: https://github.com/krishnaveti/gpt_takes_sat_replication; any further inquiries can be directed to the corresponding author.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript. Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors.

References

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review* 93, 113–132. doi: 10.1257/000282803321455188

Abdelkarim, S., Lu, D., Flores, D.-L., Jaeggi, S., and Baldi, P. (2025). Evaluating the intelligence of large language models: A comparative study using verbal and visual IQ tests. *Computers in Human Behavior: Artificial Humans* 5, 100170. doi:10.1016/j.chbah.2025.100170

Abdurahman, S., Ziabari, A. S., Moore, A. K., Bartels, D. M., and Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science* 8. doi:10.1177/25152459251325174

Anthropic (2024). Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-07-28

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*

Belasco, A. S., Rosinger, K. O., and Hearn, J. C. (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis* 37, 206–223. doi:10.3102/0162373714537350

Bennett, C. T. (2021). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal* 59, 180–216. doi:10.3102/00028312211003526

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., et al. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation* 5, Article 3. doi:10.59863/MIQL7785

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics* 225, 200–230. doi:10.1016/j.jeconom.2020.12.001

Card, D. and Rothstein, J. (2007). Racial segregation and the black–white test score gap. *Journal of Public Economics* 91, 2158–2184. doi:10.1016/j.jpubeco.2007.03.006

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 1–45

College Board (2016). *Test Specifications for the Redesigned SAT*. Tech. rep., College Board, New York, NY. Available at: https://satsuite.collegeboard.org/media/pdf/test-specifications-redesigned-sat-1.pdf. Accessed: 2023-10-20

College Board (2019). *SAT Scoring and Equating*. Tech. rep., College Board, New York, NY. Accessed: 2025-05-05

Cunningham, S. (2021). *Causal Inference: The Mixtape* (Yale University Press). doi:10.2307/j.ctv1c29t27

Dell, M. (2024). Deep learning for economists. *Journal of Economic Literature* 62, 1281–1317. doi:10.1257/jel.20241733

Gao, Y., Bing, L., Li, P., King, I., and Lyu, M. R. (2019). Difficulty controllable question generation for reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4837–4847

Guo, H., Robin, F., and Dorans, N. (2017). Detecting item drift in large-scale testing. *Journal of Educational Measurement* 54, 265–284

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., and Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice* 43, 16–29

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The Review of Economics and Statistics* 91, 717–724. doi:10.1162/rest.91.4.717

Holland, P. W. and Dorans, N. J. (2006). Linking and equating. *Educational measurement* 4, 187–220

Ji, J., Qiu, T., Chen, B., Zhou, J., Zhang, B., Hong, D., et al. (2025). AI alignment: A contemporary survey. *ACM Computing Surveys* 58, 1–38

Kaldaras, L., Akaeze, H. O., and Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. In *Frontiers in education* (Frontiers Media SA), vol. 9, 1399377

Kang, H.-A. (2023). Sequential generalized likelihood ratio tests for online item monitoring. *Psychometrika* 88, 672–696. doi:10.1007/s11336-022-09871-9

Kiciman, E., Ness, R., Sharma, A., and Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*

Kolen, M. J. and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices* (New York: Springer), 3rd edn. doi:10.1007/978-1-4939-0317-7

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature* 61, 1281–1317. doi:10.1257/jel.20231736

Lee, Y.-H. and Lewis, C. (2021). Monitoring item performance with CUSUM statistics in continuous testing. *Journal of Educational and Behavioral Statistics* 46, 611–648. Accessed: 2026-02-12

Liu, J., Jiang, Z., Zheng, T., Han, Y., and Feng, S. (2025). Common persons design in score equating: A monte carlo investigation. *Educational and Psychological Measurement* , 00131644251380585doi:10.1177/00131644251380585. Epub ahead of print

Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement* 34, 365–369. doi:10.1177/0146621609359626

National Center for Education Statistics (2023). *NAEP Long-Term Trend Assessment Highlights 2023*. Tech. rep., U.S. Department of Education. Available at: https://www.nationsreportcard.gov/highlights/ltt/2023/. Retrieved November 21, 2024

National Center for Education Statistics (2024a). Fast facts: Long-term trends in reading and mathematics achievement. https://nces.ed.gov/fastfacts/display.asp?id=38. Accessed: 2025-05-05

National Center for Education Statistics (2024b). Long-term trend assessments: Reading and mathematics. https://nces.ed.gov/nationsreportcard/ltt/. Accessed: 2025-05-05

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., et al. (2022). Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. (2023). Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. *arXiv:2303.17003* Accessed: 2023-10-21

OpenAI (2023a). *GPT-4 Technical Report*. Tech. rep., OpenAI. Available at: https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2023-10-21

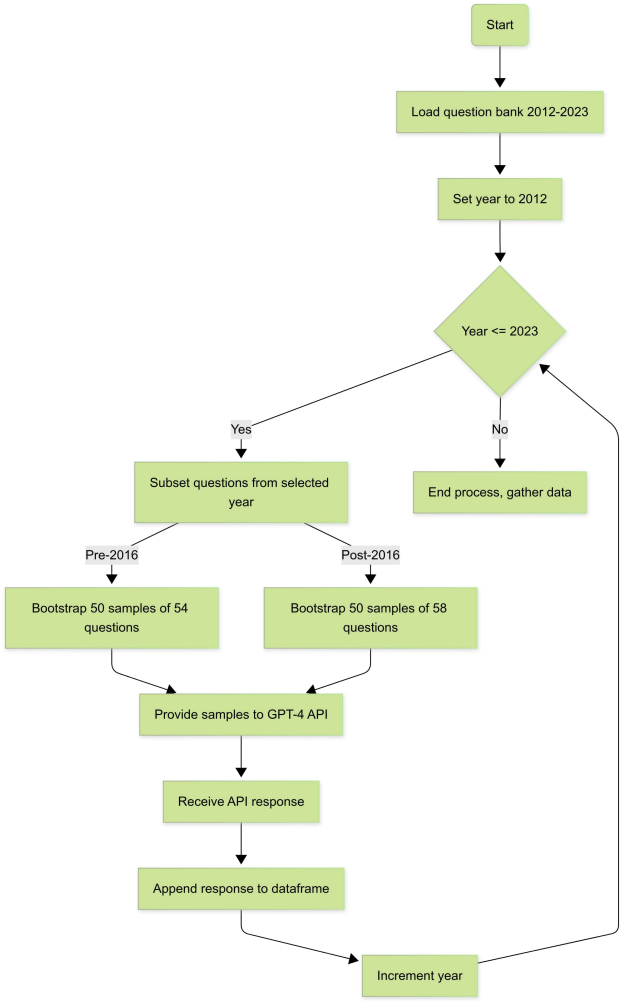
OpenAI (2023b). Gpt-4 turbo and gpt-4 models. https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4. Accessed: 2024-04-15

Reardon, S. F. (2018). The widening academic achievement gap between the rich and the poor. In *Inequality in the 21st Century*, eds. D. B. Grusky and J. Hill (Routledge), chap. 13. 1 edn., 177–189. doi:10.4324/9780429499821-29

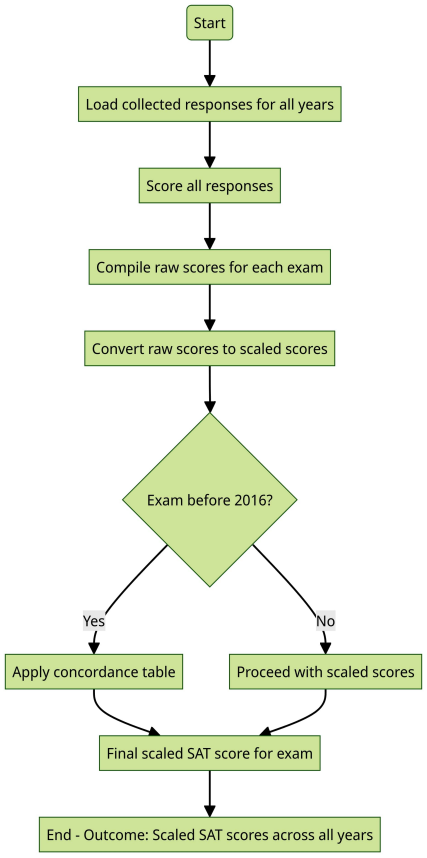
Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., et al. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence* 3, 100075. doi:10.1016/j.caeai.2022.100075

Xue, D., Wu, H., Hong, Z., Guo, S., Gao, L., Wu, Z., et al. (2018). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems* 33, 74–82

Zhao, C., Habule, M., and Zhang, W. (2025). Large language models (llms) as research subjects: Status, opportunities and challenges. *New Ideas in Psychology* 79, 101167. doi:10.1016/j.newideapsych.2025.101167

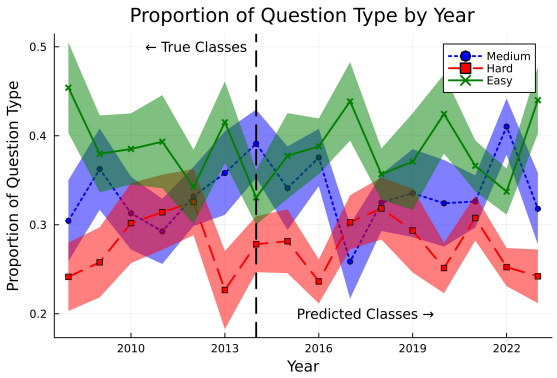


(a) GPT-4 Answering

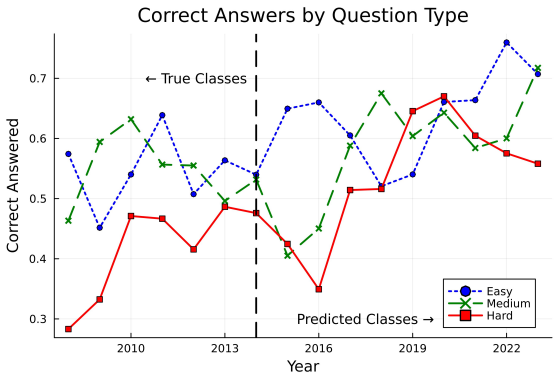


(b) GPT-4 Scoring

Figure 2: GPT-4 Data Processing Flowcharts

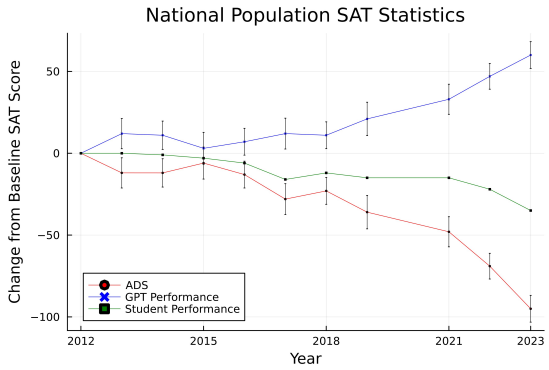


(A) SAT Difficulty Consistency

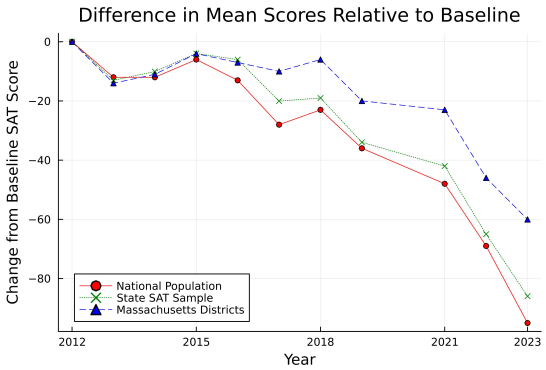


(B) Accuracy Rate by Question Difficulty

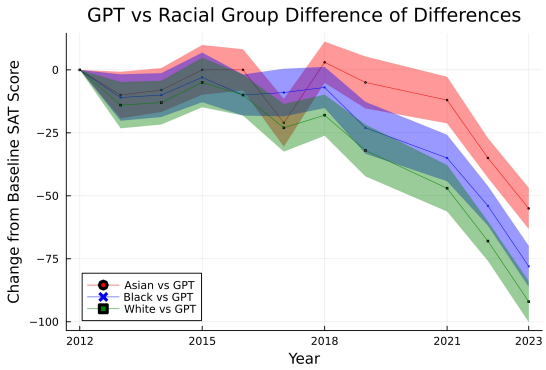
Figure 3: Question Difficulty Alignment



(A) Estimated $ADS(\hat{t}, t)$, National Population

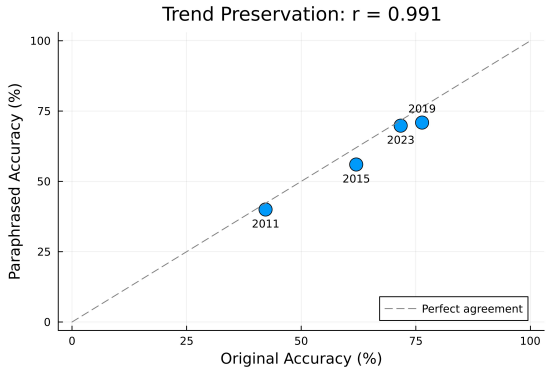


(B) Estimated $ADS(\hat{t}, t)$

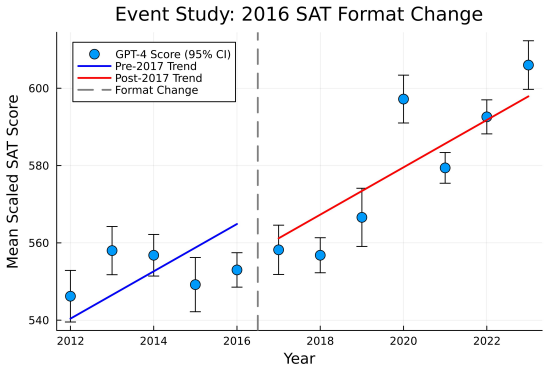


(C) Estimated $ADS(\hat{t}, t)$ for Asian, Black and White Students

Figure 6: Estimated Average Difference Scores (ADS) relative to 2012 baseline. ADS = (GPT-4 change from baseline) − (Student change from baseline), measuring difficulty-adjusted student performance change. Panel (A) uses national population data ($n > 1.5$ million test-takers per year). Panel (B) compares national population, state unweighted average ($n = 51$ states), and Massachusetts district panel ($n = 228$ districts). Panel (C) shows heterogeneity by race. Error bars represent 95% confidence intervals from bootstrap simulation (100,000 draws from t-distributions).



(A) Paraphrasing Validation



(B) Event Study: 2016 Format Change

Figure 7: Robustness checks. Panel (A) shows the correlation between GPT-4 accuracy on original vs. paraphrased questions ($r = 0.991$), indicating temporal trends are preserved when surface text is altered. Panel (B) shows an event study at the 2016 SAT redesign; the jump coefficient is -9.82 (SE=11.96) and is statistically insignificant ($p = 0.433$), confirming no discrete discontinuity at the format change.

Appendix

Appendix A: Concordance Tables

The concordance tables provided below come directly from the College Board. We utilize these concordance tables in our paper for 2 types of score conversions: converting raw to scaled score and converting old to new score. The table below shows the conversion tables utilized by our study.

Appendix B: Data Sources

After collecting the PDF files, we transcribed each SAT exam PDF into a structured comma-separated values (CSV) file. MCQs present a set of possible answers, requiring the examinee to select the most appropriate option. Answer type allowed a one-line input from the examinee.

The College Board provides yearly reports with the population level SAT performance by the cohort of high school students taking the SAT exam. The data from these reports include average scores, total test takers and standard deviation in the scores for the mathematics test and the language and writing test. The concordance table provided by the College board requires rounding the average SAT score for exams in the *pre* period to the nearest multiple of 10 before it can be mapped to the average SAT score based on the concordance table. Due to this, the direct conversion of SAT scores using the concordance table has certain limitations. For example, if an average SAT score is 514 in the year 2009 and the score is 515 in the year 2010, 514 would be rounded to 510 and 515 would be rounded to 520, before the concordance table can be used. To avoid this problem, we used a simple linear regression model to regress average SAT scores before and after the conversion in the concordance table and then linearly interpolate the average SAT scores, which we call the Concordance SAT scores.

Negative marking

As previously stated, our study spans two eras, 2008-2016 known as *pre* and 2017-2023 known as *post* period. During *pre* period, the exams implemented negative marking, unlike the 2017-2023 period. In analyzing the GPT models' performance, we opted not to apply negative scoring for incorrect responses when converting from raw score to scaled score. This decision might initially suggest an inflated assessment of GPT performance in the *pre* period, as measured by scaled SAT scores. However, as our forthcoming sections demonstrate, incorporating negative marking for this

Table E.2: Concordance Tables for raw-to-scaled and pre-to-post score conversions

Post		Pre		Pre	Post
Raw Score	Scaled Score	Raw Score	Scaled Score	Scaled Score: Old	Scaled Score: New
58	800	54	800	200	200
57	790	53	790	210	220
56	780	52	760	220	230
55	760	51	740	230	250
54	750	50	720	240	260
53	740	49	710	250	280
52	730	48	700	260	300
51	710	47	690	270	310
50	700	46	680	280	330
49	690	45	670	290	340
48	680	44	660	300	350
47	670	43	650	310	360
46	670	42	640	320	360
45	660	41	640	330	370
44	650	40	630	340	380
43	640	39	620	350	390
42	630	38	610	360	400
41	620	37	600	370	410
40	610	36	590	380	420
39	600	35	590	390	430
38	600	34	580	400	440
37	590	33	570	410	450
36	580	32	560	420	460
35	570	31	550	430	470
34	560	30	540	440	480
33	560	29	540	450	490
32	550	28	530	460	500
31	540	27	520	470	510
30	530	26	510	480	510

Table E.3: Concordance Tables for raw-to-scaled and pre-to-post score conversions (continued)

Post		Pre		Pre	Post
Raw Score	Scaled Score	Raw Score	Scaled Score	Scaled Score: Old	Scaled Score: New
29	520	25	500	490	520
28	520	24	490	500	530
27	510	23	480	510	540
26	500	22	480	520	550
25	490	21	470	530	560
24	480	20	460	540	570
23	480	19	450	550	570
22	470	18	440	560	580
21	460	17	430	570	590
20	450	16	420	580	600
19	440	15	420	590	610
18	430	14	410	600	620
17	420	13	400	610	630
16	410	12	390	620	640
15	390	11	380	630	650
14	380	10	370	640	660
13	370	9	360	650	670
12	360	8	350	660	690
11	340	7	330	670	700
10	330	6	320	680	710
9	320	5	310	690	720
8	310	4	290	700	730
7	290	3	280	710	740
6	280	2	260	720	750
5	260	1	240	730	760
4	240	0	220	740	760
3	230	-1	200	750	770
2	210	-2	200	760	780
1	200			770	780
0	200			780	790
				790	800
				800	800

period would likely have amplified the performance discrepancy between the two periods. Therefore, our analysis likely presents a conservative estimate, potentially downplaying the actual performance gap.

Syllabus change

Although SAT math section did not experience a major change in structure and syllabus as compared to the verbal section, we must emphasize certain differences in exam content between the two periods. Before 2016, the key focus areas were arithmetic, numbers and operations, algebra, functions, geometry and data analysis. Starting 2016, SAT exam expanded the range of topics further and included questions related to trigonometry and complex numbers. Also, there was a greater focus on data analysis, graphs and word problems and less emphasis on geometry-related questions. Furthermore, starting 2016, students started received sub-scores for sections labeled as *Heart of Algebra*, *Passport to Advanced math* and *Problem Solving and Data Analysis*. There were no such subsections before 2016. Just by looking at the syllabus, one could argue that the math section became harder than before as it covered the number of topics covered increased, along with their difficulty level.

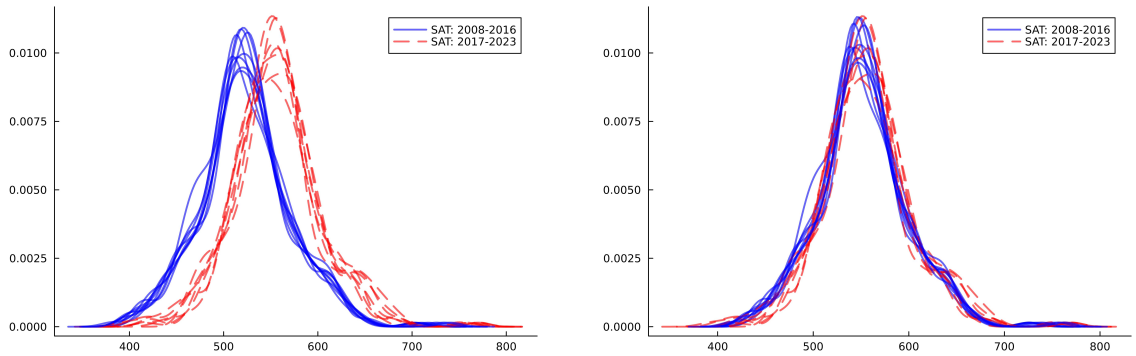
Data Sources

These reports provide comprehensive insights into the SAT performance of students on a state-by-state basis, enabling a detailed examination of trends and patterns in SAT scores across the United States. However, the availability of state-level SAT data prior to 2016 posed a unique challenge. To procure this historical data for the *pre* period (2008-2016), we used the Internet Archive and the National Center for Education Statistics. This digital archive, renowned for its extensive collection of web pages archived over time, proved instrumental in retrieving past state-level SAT reports not readily accessible on the College Board's current website. Utilizing this, we systematically sourced and compiled state-level SAT score reports for each year from 2008 to 2016. This approach of combining current and archived data sources ensures a comprehensive and continuous dataset spanning the entire duration of our study period.

Data Cleaning and Recentering

The distribution of the average scores for these school districts is provided in Figures 8a and 8b, which show the distribution of average scores before and after the recen-

tering using the concordance tables from the *pre* period to match the *post* format change period.



(a) Before Conversion

(b) After Conversion

Figure 8: SAT Score Distribution for School Districts in Massachusetts

Appendix C: Prompting

OpenAI provides access to their GPT series models through Application Programming Interface (API) endpoint. In this study, we used the most advanced model available GPT-4 . This model has been bench marked by OpenAI to be performing at the 89th percentile on SAT mathematics under certain prompting conditions as seen in OpenAI (2023a). For the purposes of this study, we employed a prompting strategy with explicit system level instructions to return the appropriate character letter corresponding to the correct answer having analyzed the question provided. Prompting strategies can affect the performance of the model as described by Nunes et al. (2023). Since, we are concerned with change in the LLM agent performance over time with reference to baseline, any idiosyncratic LLM effects and prompt effects cancel out. We employed a zero-shot prompt, by which, no examples solutions were provided to the model to facilitate answering the question. Further, we did not allow for chain-of-thought. The model was proscribed from sequentially reasons itself to the appropriate answer. The question was provided through the prompt and the model was asked to provide one character letter output for multiple choice questions and the appropriate numerical or equation output for the answer type questions. The prompting strategy is identical over all periods, ensuring there is no bias through the prompt. Since all prompts are independent API calls to the model, the model has no memory of previous questions. The model therefore is tasked to

answer each question independent of any other question or reference to the period from which the question is gathered. This ensured the comprehensive evaluation of the difficulty of the questions from appropriate periods.

The important parameters concerning this process are the temperature of the model and maximum token output limit. The temperature of the model governs the randomness of the output, the parameter when set to 0 provides a consistent and unwavering output from the LLM agent. This ensures the output from the LLM agent is deterministic and repetitive by always selecting the most probable token given the previous token. As the temperature is increased to higher levels, the agent has additional creative license to provide more abstract outputs. Since all LLM agents are predictive engines for the next token, higher temperature increases the probability of less likely token being selected as the output. For the purposes of our experiment and not to induce bias through the prompt temperature, we set it to 0. The maximum token output parameter limits the LLM agent's output. Since the agent in our experiment is required to choose the correct multiple choice option or provide straightforward numerical output, we limit the agent to 5 tokens, which is roughly 15-20 characters.

Our prompt template is shown in Listing 1, it encapsulates the strategy employed throughout the experiment. The LLM agent operates as an assistant to the user and to illicit a response from this assistant, two level of prompts are used. The system-level prompt is a high level instruction that the assistant is expected to follow verboten, this gives necessary context to the LLM agent and appropriately modulates its responses. For example, the LLM agent can be required to respond as a Shakespearean character through the system-level instructions. This is less useful in our case but provides the necessary platform to provide broad instructions to the agent about the task it is undertaking. The user-level prompt then provides the SAT question to the agent, the agent then responds to the output based on the temperature and maximum token output limit. Once these prompts and model parameters are held constant, we loop through the bootstrapped exams from each year as shown in Figure 2b. This ensures our agent is neutrally evaluating the SAT exam in the selected year while all other parameters are held constant.

Listing 1: SAT math Exam Instructions for GPT-4

```
@System-level prompt@
You are taking an SAT math exam which include multiple
   choice and answer type questions.
Determine the correct answer.
Choose only ONE 'character letter' response output
```

```
    corresponding to the correct answer from the options
    provided for multiple choice type.
Provide the appropriate numerical answer as required for
    the answer type question without any units of
    measurement.

@User-level prompt (for multiple-choice questions)@
You are provided with an SAT question enclosed in triple
    backticks, followed by multiple choice options.
``` <Question>
    <Options> ```
Please identify and return ONLY ONE letter character
    corresponding to the correct option.
Your response output should only be ONE character letter
    .

@User-level prompt (for numerical answer questions)@
You are provided with an SAT question of numerical type,
    enclosed in triple backticks. Please determine and
    return the correct numerical value or mathematical
    expression.
``` <Question> ```
WARNING: Do not provide any explanations, calculations,
    units of measurement, or additional outputs.
```

Appendix D: Other LLMs

To ensure the robustness of our findings and to validate that our results are not subject to the specific LLM used, we conducted additional tests using different versions of the LLM. We utilized the GPT-4 April update and another LLM, Claude 3.5 Sonnet, for this purpose. GPT-4 Turbo received an update on April 9th, 2024. This update by OpenAI majorly improved the model performance (OpenAI, 2023b). The results in the main text are from GPT-4 January update, and we provide the results from GPT-4 April update. The results show a similar trend in the decline of student performance over time. The decline in SAT scores is 113 points at the national level in 2023 compared to 2008. The decline in SAT scores is 72 points at the Massachusetts level in 2023 compared to 2008. Additionally, it can be noted through Figure 9 that the performance of GPT-4 April tracks the GPT-4 January in

evaluating underling SAT. The results are consistent with the main text and show a decline in SAT math difficulty over time.

Next, we utilized Claude 3.5 Sonnet, a different LLM, to evaluate the SAT exams. Claude 3.5 Sonnet is a LLM developed by Anthropic, a U.S based artifical intelligence (AI) company. We used this model as model benchmarking performed by Anthropic show similar performance results compared to OpenAI's GPT-40 model (Anthropic, 2024). The results from Claude 3.5 Sonnet are also provided in Figure 9. In the figure, we can clearly see that the initial and end point performance of Claude 3.5 Sonnet on the SAT exams is consistent with the performance of GPT-4 January and GPT-4 April update. The overall trend of declining SAT scores from 2008 to 2023 was replicated by these runs, further strengthening our claim. This robustness test demonstrates that our findings are consistent across different LLMs, provided that the intelligence demonstrated by LLMs is comparable, and are not an artifact of the specific LLM used in the initial analysis.

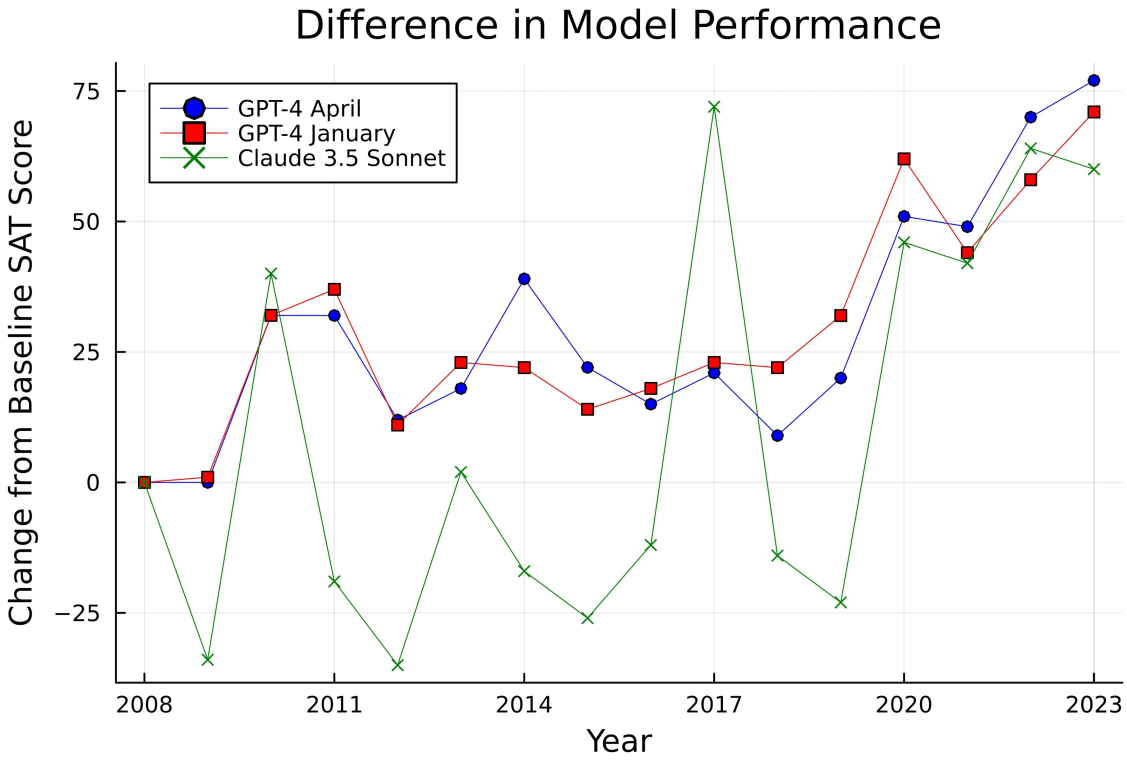


Figure 9: Change in Performance of GPT-4 April and GPT-4 January

Appendix E: Additional Tables and Figures

Table E.4: Yearly Performance of LLM Agent by Question Type

Year	All Questions		Multiple Choice only		Answer Type only	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
2008	0.469	(0.008)	0.495	(0.009)	0.378	(0.018)
2009	0.473	(0.007)	0.459	(0.008)	0.523	(0.013)
2010	0.548	(0.006)	0.525	(0.007)	0.630	(0.013)
2011	0.560	(0.008)	0.555	(0.009)	0.578	(0.011)
2012	0.494	(0.008)	0.493	(0.009)	0.500	(0.014)
2013	0.523	(0.008)	0.546	(0.009)	0.445	(0.013)
2014	0.519	(0.006)	0.543	(0.008)	0.435	(0.013)
2015	0.502	(0.008)	0.517	(0.010)	0.448	(0.015)
2016	0.508	(0.005)	0.516	(0.006)	0.478	(0.007)
2017	0.574	(0.007)	0.603	(0.008)	0.474	(0.015)
2018	0.569	(0.005)	0.592	(0.006)	0.492	(0.009)
2019	0.592	(0.008)	0.630	(0.008)	0.462	(0.018)
2020	0.657	(0.007)	0.654	(0.008)	0.669	(0.015)
2021	0.619	(0.004)	0.578	(0.006)	0.762	(0.005)
2022	0.648	(0.005)	0.672	(0.006)	0.566	(0.011)
2023	0.675	(0.007)	0.699	(0.007)	0.592	(0.018)

Notes: This table reports the average proportion of questions correctly answered by the LLM agent each year. The estimates and standard errors (S.E.) for the proportion of questions correctly answered by question type are provided. The estimates represent the ratio of correctly answered questions to the total number of questions in that category of question type.

Table E.5: Asian, Black and White Student SAT Score Changes

Year	Asian		Black		White	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
2009	3.595	(4.764)	-1.326	(4.74)	-2.493	(4.73)
2010	-22.978	(4.255)	-29.926	(4.204)	-33.027	(4.203)
2011	-24.386	(4.797)	-36.367	(4.818)	-39.425	(4.819)
2012	1.832	(5.099)	-9.17	(5.039)	-12.268	(5.001)
2013	-7.993	(4.945)	-19.935	(4.881)	-26.079	(4.899)
2014	-5.761	(4.599)	-18.729	(4.609)	-24.826	(4.623)
2015	1.808	(5.195)	-12.163	(5.177)	-17.219	(5.179)
2016	1.993	(4.331)	-18.94	(4.295)	-22.032	(4.329)
2017	-19.174	(4.988)	-18.104	(4.88)	-35.219	(4.969)
2018	5.208	(4.371)	-15.736	(4.318)	-29.852	(4.351)
2019	-2.59	(5.358)	-31.52	(5.423)	-43.61	(5.454)
2020	-38.206	(4.904)	-65.152	(4.859)	-80.246	(4.873)
2021	-10.403	(4.171)	-44.337	(4.179)	-59.447	(4.248)
2022	-32.568	(4.31)	-62.526	(4.347)	-79.617	(4.302)
2023	-52.988	(4.924)	-86.969	(4.898)	-103.99	(4.892)

Table E.6: Male and Female Student SAT Score Changes

Year	Male		Female	
	Estimate	S.E.	Estimate	S.E.
2009	-0.385	(4.766)	-2.353	(4.725)
2010	-30.991	(4.248)	-32.007	(4.207)
2011	-39.401	(4.847)	-37.417	(4.84)
2012	-12.23	(5.051)	-12.199	(5.055)
2013	-24.989	(4.923)	-23.996	(4.891)
2014	-24.796	(4.688)	-22.831	(4.581)
2015	-20.182	(5.176)	-18.197	(5.152)
2016	-26.988	(4.35)	-23.984	(4.292)
2017	-46.147	(4.863)	-36.227	(4.925)
2018	-40.797	(4.357)	-28.773	(4.299)
2019	-55.57	(5.359)	-41.587	(5.356)
2020	-92.155	(4.91)	-75.156	(4.913)
2021	-68.379	(4.234)	-54.405	(4.151)
2022	-88.565	(4.349)	-74.586	(4.293)
2023	-117.006	(4.897)	-100.039	(4.876)

Table E.7: Yearly National and Massachusetts SAT Score Changes after removing post-COVID years

Year	National		Massachusetts	
	Estimate	S.E.	Estimate	S.E.
2009	-1.831	(5.495)	1.052	(3.208)
2010	-32.471	(5.495)	-27.776	(3.208)
2011	-39.792	(5.495)	-34.448	(3.208)
2012	-13.357	(5.495)	-6.112	(3.208)
2013	-25.980	(5.495)	-20.404	(3.208)
2014	-23.271	(5.495)	-17.173	(3.208)
2015	-17.063	(5.495)	-10.441	(3.208)
2016	-18.784	(5.495)	-13.114	(3.208)
2017	-33.416	(5.495)	-15.897	(3.208)
2018	-31.643	(5.495)	-11.861	(3.208)
2019	-46.522	(5.495)	-26.236	(3.208)
2020	-83.749	(5.495)	-54.599	(3.208)

Notes: This table reports the decline after removing post-COVID years. The national sample represented by the unweighted state SAT data.

Appendix G: Question Embeddings and Difficulty Alignment

The difficulty of the questions is predicted using a Random Forest Classifier trained on the question embeddings and question location within the section as features. The Table E.8 provides the classifier accuracy for the test sample. The standard error is calculated using the proportion of predicted difficulty rating for questions in the bootstrapped SATs. It does not account for the uncertainty in the classifier for the unseen data between 2015 and 2023.

Table E.8: Confusion Matrix: Classification Accuracy - Test Sample

Predicted/Ground Truth	**Easy**	**Medium**	**Hard**
Easy	130	15	1
Medium	17	61	23
Hard	1	4	36