

How do LLMs take the SAT and ACT?

**Sophie
Juco**
smj490@nyu.edu

**Anagha
Radhakrishna Palandye**
ap8884@nyu.edu

**Simone
Rittenhouse**
sgr344@nyu.edu

**Taruni
Nugooru**
tn2463@nyu.edu

Abstract

Standardized tests are used in college admissions as a measure of scholastic aptitude, but it remains unclear whether test-takers rely on reason or learn test-specific heuristics for good performance. To uncover potential heuristics for common U.S. standardized exams, we fine-tune GPT-2 XL, a large language model (LLM), on two standardized assessments: official SAT and ACT reading comprehension questions and RACE-H – a similar high school-level multiple-choice question reading comprehension dataset. Model performance is compared using a series of interpretability experiments. This study contributes a novel dataset of SAT and ACT reading comprehension questions as well as insights on the strategies learned by LLMs for this task. We find that the fine-tuned SAT/ACT model obtains higher performance than both the base and RACE-H models as well as demonstrated learning of test-specific features.

1 Introduction

The Scholastic Aptitude Test (SAT) is a multiple-choice question exam designed to measure skills in reading, writing, and math (CollegeBoard, 2025). It, alongside its counterpart the American College Testing (ACT) exam (ACT, 2025), are commonly used in the United States to evaluate applicants during the college admissions process. While in recent years there has been less focus on these exams in college admissions, they remain an important milestone of many American students.

Much work has been done to assess the validity of such exams as predictors of college success (Rothstein, 2004), but questions remain as to whether the SAT and ACT evaluate reasoning or some other factor (i.e. socioeconomic status, academic support, etc.). It is therefore crucial to clarify what standardized exams are testing.

We employ large language models (LLMs) to offer insights as to whether students can game their way through these exams. To do this, we compare

the performance of two GPT-2 XL models: one fine-tuned on SAT and ACT reading comprehension questions and another fine-tuned on a comparable reading comprehension multiple-choice dataset (RACE-H). This comparison focuses on interpretability to reveal test specific features being used by the model. We hypothesize:

- The SAT/ACT model will learn test specific features that increase performance from the baseline pretrained GPT-2 XL model.
- The SAT/ACT model will deploy different heuristics than the model fine-tuned on a comparison dataset.

We find that the SAT/ACT exam’s unique vocabulary and syntax can be learned by the model for improved performance over both the base and RACE-H models in baseline accuracy and chain-of-thought rationale generation. The SAT/ACT model shows a higher sensitivity to token-level changes as well as above-chance performance when provided with the question and answer choices alone. Taken as a whole, these findings indicate that students may be able to boost their performance through surface-level, syntax-based strategies.

2 Related Work

Previous work has used LLMs to evaluate the quality of standardized exams. Suresh and Rawat (2024) applied GPT-4 to SAT math questions in order to track changes in exam difficulty over time. While this offers a method of assessing exam difficulty, it does not inherently clarify how the model learns to complete the exam. McCoy et al. (2019) found that statistical natural language inference models like LLMs can score well on evaluation benchmarks simply by learning a set of heuristics that are effective for common example types. This style of learning mimics how students are taught tricks to more effectively process standardized exams. Finding the heuristics used by LLMs on the

SAT and ACT can offer effective test-taking strategies.

Despite this work, the SAT/ACT’s potential as a reading comprehension evaluation benchmark has been overlooked. LLM benchmarks like AGIEval (Zhong et al., 2024) use as few as 206 reading and writing SAT questions. Our work offers a novel, comprehensive dataset of official SAT and ACT reading comprehension questions that can be used in future work for further LLM evaluation.

3 Data

We created a dataset of 1,315 SAT and ACT reading comprehension questions scraped from official practice tests and questions. Of these, 1,194 are from the SAT and 121 are from the ACT. The difference in number between the two exams is due to the availability of official questions provided publicly by the two testing companies. Each entry contains a relevant reading passage, question, four answer options, and a final answer key as shown in the following example:

prompt: At two weeks old, wolves can smell but cannot yet see or hear. Domesticated dogs, _____ can see, hear, and smell by the end of two weeks.
question: Which choice completes the text with the most logical transition?
mc_a: in other words,
mc_b: for instance,
mc_c: by contrast,
mc_d: accordingly,
answer: C

For our comparison dataset, we selected the high school level questions of the RACE dataset (Lai et al., 2017). This dataset is designed to test the reading comprehension level of non-English speaking students, and has been selected due to its similar question styles and exam format. The RACE-H dataset contains 69,394 questions.

The questions are comparable to the SAT and ACT exam questions in both format and length (SAT/ACT: 122.46 average prompt tokens, 19.20 average question tokens; RACE-H: 355.09 average prompt tokens, 11.40 average question tokens). Additionally, the Flesch Reading Ease metric is 46.18 for the SAT and ACT questions and 64.52 for RACE-H. This makes the SAT/ACT a slightly more difficult but still comparable dataset.

4 Models

The study uses GPT-2 XL, which was chosen as the weights are publicly available, it is a manageable size given our resource constraints, and because GPT variants have been shown to perform well in the current setting (Suresh and Rawat, 2024; Sun et al., 2019). GPT-2 XL is a 1.5 billion parameter transformer-based LLM released by OpenAI (Radford et al., 2019).

We define the model’s answer to a question q as:

$$\arg \max_i \mathbb{P}[a_i | pq], \quad i \in \{1, 2, 3, 4\}$$

where a_i represent each of the four possible answer choices and p represents the corresponding question context (i.e., the reading passage). This is equivalent to taking the minimal cross entropy loss between the concatenated texts pqa_1, \dots, pqa_4 . We use an MC1 pipeline similar to that of Lin et al. (2025), particularly in formatting the input text. Figure 8 outlines the default prompt format.

Inputs longer than the model’s context window size of 1024 tokens were truncated from the front of the input in order to preserve the question and answer choices. However, since the number of inputs over 1024 tokens accounts for <2% of the total data, we dropped truncated questions for our final test dataset.

Inference was run on a single 16 GB T4 GPU, and fine-tuning was conducted on a 40GB A100 GPU.

4.1 Fine-Tuning

To accomplish fine-tuning within our resource constraints, low rank adaptation (LoRA) was used as detailed in Hu et al. (2021) for memory efficiency. Specifically, an attention dimension of 16 was used with scaling $\alpha = 16$ and a dropout probability of 0.05 for the LoRA layers. Only the attention, fully connected, and embedding layers were trainable during the fine-tuning process. This equated to training 1.30% of the total 1.5 billion parameters.

As a form of data augmentation, we implemented the Highlighting (HL) and Self-Assessment (SA) methods detailed in Sun et al. (2019) – which have been shown to improve performance in the current setting. The HL strategy adds trainable embeddings to the tokens deemed relevant to the question and answer choices. In the SA strategy, questions and their relevant answer options are randomly generated from reference documents to create novel practice questions. The practice questions

are created by randomly sampling sentences from the reference document, and the associated “correct” answer options are created by concatenating randomly sampled non-overlapping spans of text from the practice questions. After this process, the other three “incorrect” options are created by randomly replacing spans in the correct option with spans randomly selected from the reference document.

The training sets for both test datasets were fixed to an equal length (919 training; 131 validation). This was done both to ensure that the two models have a comparable number of training examples and because of the large training time of using the full RACE-H dataset (>24 hours on 2 V100 GPUs).

Using only the augmented training data, hyperparameter tuning was conducted across values of learning rate (1e-3, 1e-4, 5e-5) and the number of training epochs (2, 3, 5) with a loss defined as the mean cross entropy loss between ground truth answers and the model’s predictions. We find that the best parameters for both datasets are a learning rate of 1e-3 and 5 training epochs. After this initial training, the best model and hyperparameter values are used to further train on the original, unaltered training and validation sets.

5 Experiments

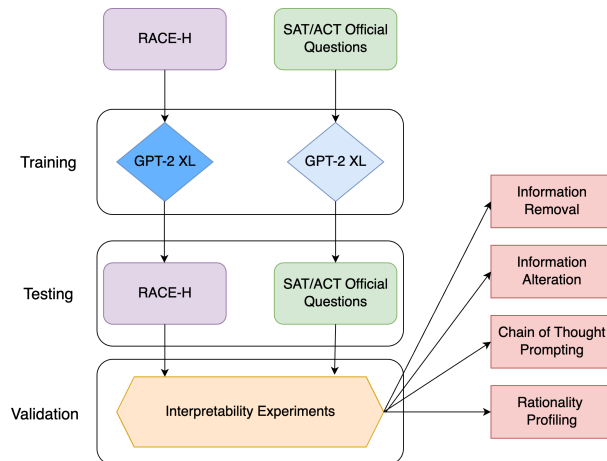


Figure 1: Overview of the experimental pipeline outlining the fine-tuning process, testing datasets, and interpretability experiments.

To assess the performance of these fine-tuned models, we conducted several interpretability assessments with the goal of assessing whether the models learn specific test strategies, what those strategies are, and how they may differ across the two exams.

5.1 Prompt Information Removal

For both the SAT/ACT and RACE-H datasets, we removed the passage from the prompt in the test set, leaving only the question and answer choices. We then re-evaluated both models and compared performance against that of the unaltered test set. The amount of decrease in accuracy after removing the passage can be interpreted as the contribution of the passage to the model’s performance. Performance higher than expected by chance on the altered test set indicates the models learned some latent pattern within the question-answer choice pairs to select the correct answer.

5.2 Prompt Information Alteration

To see the model’s reliance on test vocabulary in both the SAT/ACT and RACE-H datasets, we altered the the input passage, question, and answer choices within the test set and assessed changes in performance. In this experiment, we injected noise into the test datasets by randomly replacing words in the prompt with their synonyms. This was done by replacing words with a probability of 0.3 with a randomly chosen synonym provided by NLTK’s wordnet synset corpus. This effectively maintained the semantics of the prompt while diversifying the vocabulary and sentence structure used in the prompting.

Additionally, we executed another strategy of noise injection by randomly appending an irrelevant sentence to the passage. This was done by taking a random sentence from the previous question’s passage and appending it to the end of the current passage. Ideally, the model’s performance on both the altered and unaltered test sets should be similar, as we would expect the model to be able to ignore an irrelevant sentence. However, if the performance is worse, this would indicate that the passage as a whole is a key factor in the model’s answer.

5.3 Chain-of-Thought Prompting (CoT)

For a subset of the SAT/ACT data, we used CoT prompting to extract the model’s answering rationale – a strategy which has been shown to aid in the interpretability of LLMs (Singh et al., 2024). Specifically, we prompted for the model’s reasoning across outlier questions – answers with high confidence that are correct and answers with high confidence that are incorrect. These outlier cases enable us to understand why the model chose its an-

swer and what (in the correct case) keyed the model into the correct answer and what (in the incorrect case) confused the model.

To implement this, we adopted a one-shot and two-shot prompting approach. Each prompt included 1–2 examples containing the passage, question, answer choices, and a rationale detailing why the selected answer is correct. SAT/ACT rationales are adapted from official exam answer keys, and RACE-H rationales are generated using Grok3 and manually vetted. After the examples, the model receives a new question with no rationale present that ends with the phrase “step-by-step reasoning” to cue the model to generate its own explanation. For instance:

Step-by-step reasoning: Choice D is the best answer because in context, sharing “similarities with” means having some resemblance to.

Model responses were tested across diverse question types from both datasets. We evaluated CoT reasoning performance using two automatic metrics: BERTScore-F1 and ROUGE-1. In addition to these automatic metrics, we performed human evaluation by rating the coherence and contextual relevance of generated rationales. The scoring scheme is: 0: Completely irrelevant to the passage, question, or task, 1: One or two passage-related words/phrases, no reasoning, 2: Several passage-related words/phrases, no coherent reasoning, 3: Basic reasoning with one logical step tied to the passage, 4: Clear reasoning transition words and passage references, minor gaps possible, 5: Thorough, accurate reasoning fully addressing the question with precise passage citations.

5.4 Rationality Profiling

We conducted rationality profiling to analyze GPT-2 XL’s internal reasoning behavior, both before and after fine-tuning. The model is evaluated using a suite of interpretability metrics. Reasoning Score, defined between 0 and 1, quantifies robustness to token-level ablations by measuring changes in predictions when randomly selected input tokens (up to 20 per example) are masked. Formally, it is the total number of changed predictions (relative to the non-ablated baseline) divided by the number of masked tokens. Heuristic Dependence is the proportion of answer choices that share language with

the question. This is a model-agnostic value which measures the ability of a model to learn answer-specific information from the question itself. Context Fidelity is a binary metric that is 1 when model predictions do not match between unaltered inputs and passage-free inputs and 0 otherwise. It offers a more fine-grained view into the results of our information removal experiment.

We also compute token-wise SHAP attributions (max 20 tokens) to quantify how much each token contributes to the model’s prediction. Higher SHAP scores indicate stronger alignment with semantically meaningful evidence.

6 Results

Table 1 shows the results of the baseline and fine-tuning experiments. From these results, we see that the base GPT-2 XL model performs similarly on both datasets with low baseline accuracies.

The results of the fine-tuning show improved performance by over 30% for the SAT/ACT dataset and about 11% improved performance for the RACE-H dataset. This indicates that the training patterns of the SAT/ACT data may be easier to learn or more advantageous on the test data than those in RACE-H.

Information Removal. As shown in Table 1, we see a significant drop in performance on the information removal test set relative to the unaltered test data, with a larger gap for the higher performing fine-tuned models. We can also see that, while all of the models have an accuracy near the 25% expected by chance on the information removal test set, the fine-tuned models achieve higher performance than the base model. This indicates that the models may have learned some information about the question and answer choices that help performance even when no question context is present – particularly the SAT/ACT model which achieves the highest accuracy of 29%.

Information Alteration - Synonyms. The information alteration interpretability experiment shows worse performance on both datasets for both the base and fine-tuned models. These results indicate that the specific language and structure of the exams is important in how the model determines the correct answer. Interestingly, we see slightly lower performance compared with information removal, suggesting that the language of the exams is a greater factor in the model’s predictions than

Model	Dataset	Baseline	IR	IA - Syns	IA - Sent	CoT	
						BERTScoreF1	ROUGE-1
Base	SAT/ACT	0.307	0.246	0.239	0.303	0.57	0.21
	RACE-H	0.312	0.241	0.246	0.307	0.55	0.32
SAT/ACT fine-tuned	SAT/ACT	0.622	0.292	0.251	0.618	0.52	0.27
RACE-H fine-tuned	RACE-H	0.420	0.260	0.246	0.402	0.51	0.22

Table 1: Test dataset accuracies across baseline and fine-tuning experiments. IR: Information Removal; IA - Syns: Information Alteration - Synonyms; IA - Sent: Information Alteration - Sentences; CoT: Chain-of-Thought.

Model	Dataset	Avg Human Eval
SAT/ACT FT	SAT/ACT	2.45
	RACE-H	2.41
RACE-H FT	SAT/ACT	1
	RACE-H	0.66

Table 2: Human evaluation scores (Avg Human Eval) for fine-tuned (FT) models on SAT/ACT and RACE-H datasets for CoT experiments, averaged across 15 questions evaluated by four human annotators. Scores can fall in the range of [0, 5].

the content of the questions themselves.

Information Alteration - Sentences. The information alteration interpretability experiment performs comparably with baseline results. This indicates that both the base and fine-tuned models are able to ignore an irrelevant sentence; however, the slight loss in accuracy indicates the passage as a whole is a key factor to the model’s predictions.

Chain-of-Thought Prompting. As shown in Table 1, fine-tuning on SAT/ACT and RACE-H datasets reduced the BERTScore F1 scores for SAT/ACT and RACE-H, indicating a lower semantic similarity between the fine-tuned models’ CoT outputs and reference explanations. This discrepancy reflects fine-tuning’s prioritization of correct answer selection over CoT explanation quality.

Compared to the CoT responses generated by the base model, the SAT/ACT fine-tuned model and the RACE-H fine-tuned model reveal distinct differences in reasoning quality and structure. The SAT/ACT fine-tuned model is an improvement over the base model in both the coherence and relevance of its CoT explanations for SAT/ACT tasks (see Figure 9 in Appendix for an example). This is evident in its ability to produce logically structured reasoning that directly addresses task-specific requirements, such as Standard English conventions and textual inference, as seen in accurate responses to questions involving contextual analysis

(e.g., correctly identifying the motivation behind a character’s actions). In contrast, the base model frequently generated CoT outputs containing irrelevant tangents or extraneous details, contributing to its lower task accuracy.

Human evaluations shown in Table 2 corroborate this improvement, with SAT/ACT fine-tuned model achieving an average score of 2.45 on the SAT/ACT dataset, reflecting evaluators’ recognition of structured and task-relevant reasoning, tempered by occasional oversimplifications.

Conversely, the RACE-H fine-tuned model exhibited significant deficiencies in reasoning quality, heavily relying on one- or two-shot examples, which led to contextually inappropriate and incoherent CoT outputs. This over-reliance is apparent in its tendency to reproduce example-driven responses irrelevant to the task, such as referencing unrelated scenarios, resulting in a failure to engage with the complex reasoning demands of RACE-H questions. Human evaluations reflect this issue, which are lower than scores for the SAT/ACT model across both test sets. These findings highlight the efficacy of targeted fine-tuning in enhancing CoT reasoning for SAT/ACT fine-tuned model, while emphasizing the need for improved training strategies to mitigate RACE-H fine-tuned model’s dependence on example-based context, ensuring CoT outputs are both accurate and contextually grounded.

Rationality Profiling – SAT/ACT. Figure 2 compares interpretability metrics on SAT/ACT before and after fine-tuning. The average reasoning score increases from 1.59 to 2.32, with the distribution shifting noticeably rightward—suggesting more reliance on prompt vocabulary and increased instability to token-level changes.

There is also a noticeable shift to the right for context fidelity, which is reflective of the wider gap in performance between the baseline test set and

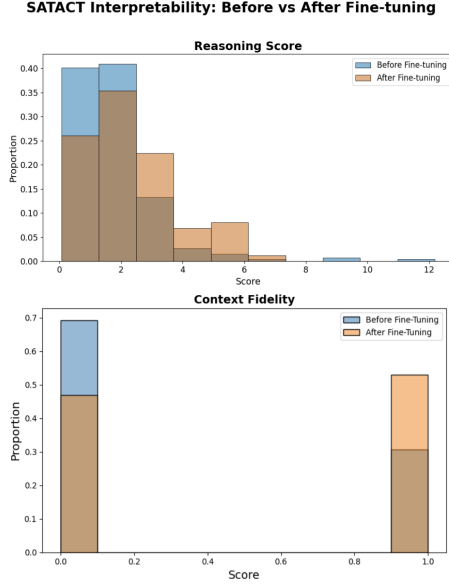


Figure 2: Interpretability metrics before and after fine-tuning on SAT/ACT.

the information removal test set post-finetuning.

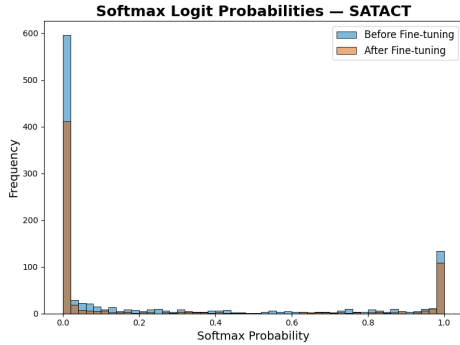


Figure 3: Softmax logit probability distribution on SAT/ACT before and after fine-tuning.

Figure 3 illustrates how the softmax probability distribution changes after fine-tuning on SAT/ACT. Prior to fine-tuning, the model assigned a wider range of intermediate probabilities, reflecting uncertainty among answer options. After fine-tuning, the distribution becomes sharply polarized, with predictions clustering near 0.0 or 1.0. This shift indicates stronger confidence in selected answers and improved ability to reject distractors.

In attention maps for the SAT/ACT fine-tuned model (Figure 4), the final layer remains heavily skewed toward the first token, but the true rollout on all the layers demonstrates stronger signal distribution, indicating deeper semantic integration across entire prompts.

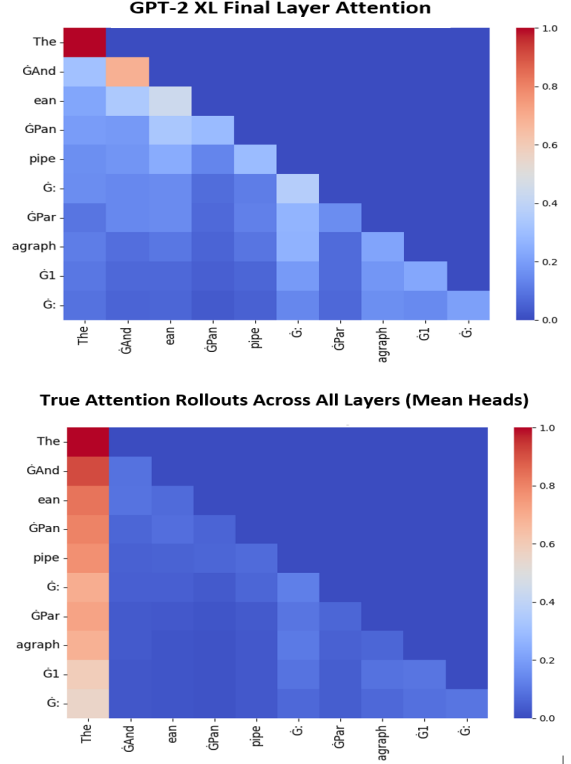


Figure 4: Top: Final Layer Attention. Bottom: True Attention Rollout on SAT/ACT (fine-tuned). Tokens beginning with \hat{G} reflect word boundaries as encoded by the GPT-2 tokenizer, where \hat{G} represents a preceding space.

We assess fine-grained interpretability on SAT/ACT using token-wise SHAP attributions (up to 20 tokens per example) and probing accuracy. SHAP scores increased from 2.21 to 2.99 post-fine-tuning, indicating greater sensitivity to semantically meaningful input features. Probing classifier accuracy remained stable around 0.62, suggesting that internal representations remained relatively linearly separable. The average heuristic dependence score is 0.542, reflecting moderate lexical overlap inherent to the dataset.

Rationality Profiling – RACE-H. Figure 5 highlights the shift in interpretability metrics on RACE-H before and after fine-tuning. The average reasoning score increases from 1.37 to 1.94, with the distribution shifting from a broader, left-skewed shape to one peaking around 2–2.5—suggesting a similar but less pronounced increase in sensitivity to token-level changes as with the SAT/ACT model.

Like the SAT/ACT model, there is also a rightward shift in context fidelity for the finetuned

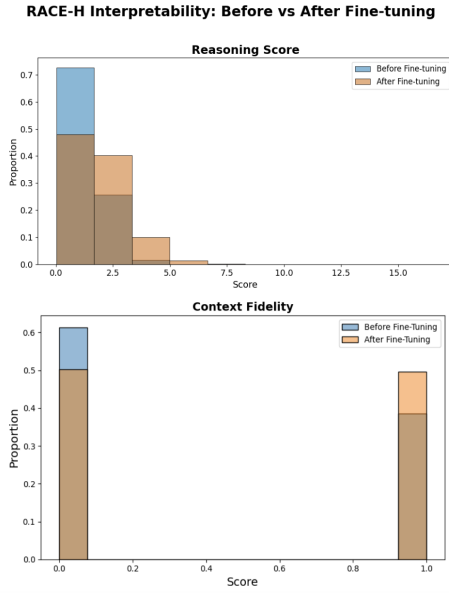


Figure 5: Interpretability metrics before and after fine-tuning on RACE-H.

model; however, the shift is less strong than with the SAT/ACT model. This suggests the RACE-H model did not learn patterns from the questions and answer choices alone to gain an edge over the base GPT-2 XL model for this task.

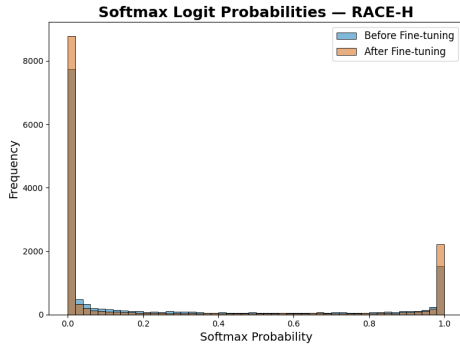


Figure 6: Softmax logit probability distribution on RACE-H before and after fine-tuning.

As seen in Figure 6, softmax probabilities become markedly more polarized after fine-tuning. While the baseline model produced a broader spread of intermediate probabilities, the fine-tuned model overwhelmingly concentrates its predictions near 0.0 and 1.0. This reflects stronger option elimination and increased confidence, suggesting improved discriminative ability between correct and distractor answers.

Figure 7 reveals residual positional bias in the final layer, though true attention rollout on all the layers is more balanced, capturing mid- and late-

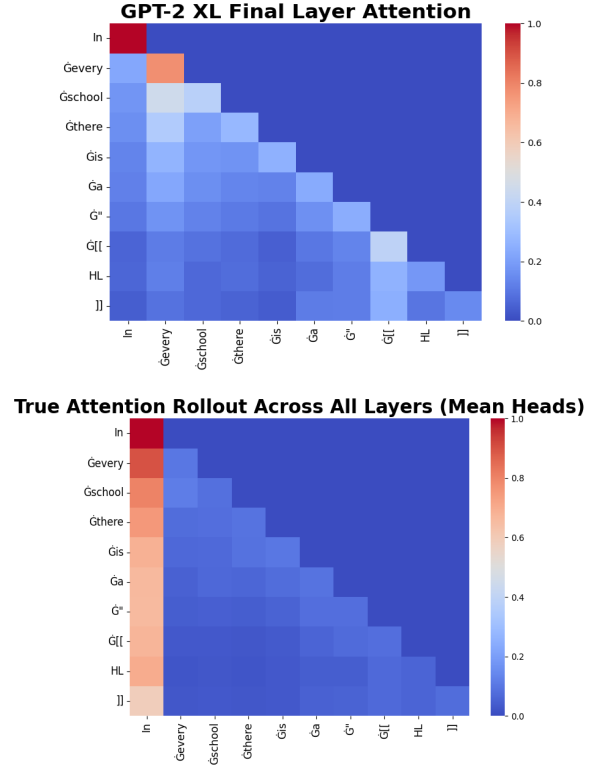


Figure 7: Top: Final Layer Attention. Bottom: True Attention Rollout on RACE-H (fine-tuned). Tokens beginning with \hat{G} reflect word boundaries as encoded by the GPT-2 tokenizer, where \hat{G} represents a preceding space.

token contributions across the input sequence.

We assessed fine-grained interpretability on RACE-H using token-wise SHAP attributions (up to 20 tokens per example), context fidelity, and probing accuracy. SHAP scores rose from 1.67 to 2.52 post-fine-tuning, indicating greater sensitivity to semantically meaningful features. Probing accuracy dropped from approximately 0.62 to 0.38, suggesting that fine-tuning produced more distributed, non-linear internal representations. The average heuristic dependence score was 0.588, reflecting moderate lexical overlap inherent to the dataset.

Overall, fine-tuning enhanced token attribution, contextual use, and internal structure, supporting more structured, evidence-driven reasoning. However, contrastive evaluation showed that model predictions remained highly sensitive to answer choice ordering both before and after fine-tuning for both the datasets.

7 Discussion

The results of our interpretability experiments indicate that the SAT/ACT model learns test spe-

cific features which increase performance from the baseline pre-trained GPT-2 XL model, confirming our first hypothesis. This is evidenced by higher accuracy on the unaltered test set, above-chance performance for information removal, and more on-task and coherent rationale generation as measured by human evaluation. Additionally, we observe an increased sensitivity to the exam’s vocabulary as shown by a higher reasoning score (indicating weakness to token-level changes) as well as chance performance in the information alteration synonym replacement experiment. Overall, the performance of the SAT/ACT fine-tuned model indicates a strong reliance on test-specific vocabulary as well as a heavy but less crucial dependence on passage and question semantic content.

Although the RACE-H model displays many of the same trends across our experiments, the smaller change in average reasoning score pre- and post- finetuning as well as the lower performance in the information removal task indicate a lowered ability to use test-specific vocabulary or phrasing when making predictions. This is indicative that the SAT/ACT model deploys a different test-taking strategy than the RACE-H model, supporting our second hypothesis. It is worth noting that the RACE-H model displays lowered performance across nearly all experiments, despite the base model performing comparatively on both test sets. This may reflect stronger or more obvious patterns in the SAT/ACT training data compared to RACE-H that can be better learned during finetuning.

In conclusion, we find that U.S. standardized exams contain inherent features that students are able to exploit for improved performance. In particular, the SAT and ACT’s narrow and specific lexicon can be learned by GPT-2 XL to improve over baseline test performance as well as performance on a comparable multiple choice exam.

8 Limitations and Future Work

Within the bounds of this study, our primary limitations were the size of our SAT/ACT dataset and the resources available to us. The SAT/ACT dataset we created is less than 2% the size of the RACE-H dataset. Additionally, ACT questions only account for about 9% of the SAT/ACT dataset. Both of these factors are due to the amount of publicly available official test and practice questions provided by the two testing companies. Ultimately,

this limits the amount of training, test, and validation data we have to fine-tune and test the models on. Larger datasets are ideal for model training and fine-tuning as the models are generally able to better fit to the training distribution and learn relevant data patterns.

The resource constraints also impacted our finetuning approach. Within the constraints of the resource sizes and time allowances, we had to scale down fine-tuning and reduce the RACE-H training and validation set sizes to match those of the SAT/ACT dataset in order to run on our provided NYU Greene resources. Future work could experiment with scaling up the number of trainable layers, the LoRA attention dimension, and the amount of training data for performance gains.

Additional work can further improve finetuning by exploring other data augmentation strategies from [Sun et al. \(2019\)](#) or similar works, as well as experimenting with different strategies of defining model predictions for the multiple-choice task. Similarly, the current training objective did not lead to improved semantic similarity between model generations and ground truth rationales in the chain-of-thought experiment as evidenced by reduced BERTScore values. To help with this task, future work could shift learning from pqa_i encoding to something better suited for rationale generation.

Another potential limitation to this study is the generalizability of our results to other LLMs. Our findings may not apply to current state-of-the-art models that have differences in architecture, pre-training, or size. We leave benchmarking the current SAT/ACT dataset across other LLMs for future work.

9 Contributions

Each interpretability method is led by a single group member with collaboration across all tasks as needed. The final contributions are:

- **Sophie Juco:** Led prompt information alteration interpretability experiments and HL and SA data augmentation strategies.
- **Anagha Radhakrishna Palandye:** Led Chain-of-Thought experiments.
- **Simone Rittenhouse:** Led the prompt information removal interpretability experiment and fine-tuning variations.
- **Taruni Nugooru:** Led rationality profiling interpretability experiments.

References

- Inc. ACT. 2025. [Act - solutions for college and career readiness](#). Official website of the ACT assessment.
- CollegeBoard. 2025. [Sat suite of assessments](#). Official website of the SAT Suite of Assessments.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *CoRR*, abs/2106.09685.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale Reading comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2025. [Truthfulqa: Measuring how models imitate human falsehoods](#). GitHub repository.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Jesse M. Rothstein. 2004. [College performance predictions and the sat](#). *Journal of Econometrics*, 121(1):297–317. Higher education (Annals issue).
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. [Rethinking interpretability in the era of large language models](#).
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. [Improving machine reading comprehension with general reading strategies](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2633–2643, Minneapolis, Minnesota. Association for Computational Linguistics.
- Vikram K Suresh and Saannidhya Rawat. 2024. Gpt takes the sat: Tracing changes in test difficulty and students’ math performance.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. [AGIEval: A human-centric benchmark for evaluating foundation models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

```
["Q: {context} {question}\nA: {a_1}",  
"Q: {context} {question}\nA: {a_2}",  
"Q: {context} {question}\nA: {a_3}",  
"Q: {context} {question}\nA: {a_4}"]
```

Figure 8: Default Prompt Format.

USER INPUT

Context: The following text is from Joan Didion's memoir *The Year of Magical Thinking*. In the text, the author discusses her home life. [I]n California we heated our houses by building fires. We built fires even on summer evenings, because the fog came in. Fires said we were home, we had drawn the circle, we were safe through the night.

Question: Which choice best describes the function of this 'Fires said we were home, we had drawn the circle, we were safe through the night.' portion in the text as a whole?

Options:

- A) It illustrates that a fire provides comfort beyond physical warmth.
- B) It summarizes the information that came before it in the text.
- C) It explains that the house remains cold even in summer.
- D) It suggests that the author feels comfortable in her home with or without a fire.

BASE MODEL OUTPUT

Step-by-step reasoning: Option A: It describes the way in which Uncle Katsihisa is 'loud' when he is not speaking.

...

Option C: It explains the feeling that the writer feels when she isn't in a home with a fire, by suggesting that she feels comfortable in her own home without one.

. . . . [

SAT/ACT FINE-TUNED MODEL OUTPUT

Step-by-step reasoning: Choice A is the best answer because it most accurately describes how the underlined portion functions in the text as a whole. The first two sentences of the text establish that in California, houses were heated by building fires year-round. The underlined portion then indicates that the fires didn't merely provide physical warmth: they also represented being "home" and feeling protected. The underlined portion thus illustrates that a fire provides comfort beyond physical warmth.

RACE-H FINE-TUNED MODEL OUTPUT

Step-by-step reasoning: Option A: It describes the way in which Uncle Katsihisa is 'loud' when he is not speaking.

Option B: It highlights the contrast between Uncle and his father, and suggests that he feels comfortable when he doesn't have a fire in his home.

Option C: It explains the feeling that the writer feels when she isn't in a home with a fire, by suggesting that she feels comfortable in her own home without one..

. . . [

Figure 9: Chain of Thought output example.