How do LLMs take the SAT and ACT?

NYU Center for Data Science

Sophie Juco, Anagha Radhakrishna Palandye, Simone Rittenhouse, Taruni Nugooru

Prof. Tal Linzen

Introduction

To uncover potential heuristics for common U.S. standardized exams, we fine-tune GPT-2 XL, on two standardized assessments: official SAT and ACT reading comprehension questions and RACE-H. Model performance is compared using a series of interpretability experiments. This study contributes a novel dataset of SAT and ACT reading comprehension questions and insights on model learning.

Data

The SAT/ACT dataset consists of 1,315 SAT and ACT reading comprehension questions scraped from official practice tests and questions – 1,194 SAT questions and 121 ACT questions.

Figure 1. Common SAT/ACT words.

Our comparison dataset was the RACE-H dataset which contains 69,394 questions. RACE-H is a multiple-choice, reading comprehension question dataset made for non-native English speakers at a high school level.

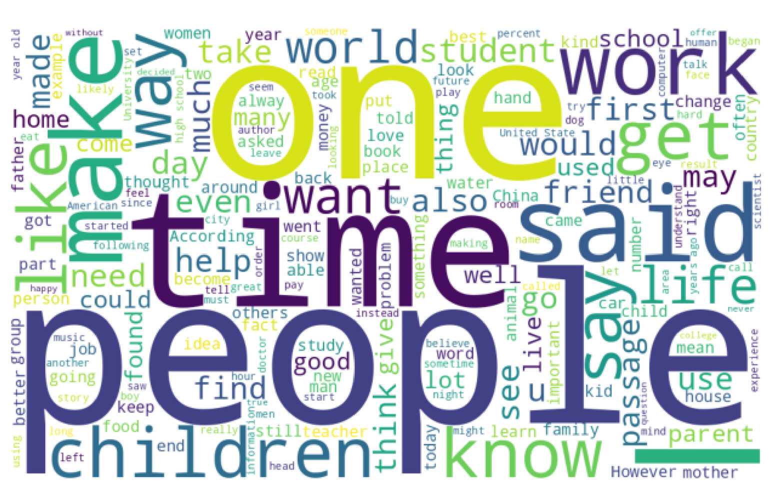

Figure 2. Common RACE-H words.

Dataset	Length (Prompt/Question)	FRE
SAT/ACT	122.46 / 19.20	46.18
RACE-H	355.09 / 11.40	64.52

Table 1. Dataset Overview Metrics.

The average number of answer choices that share language with the question is 0.542 for the SAT/ACT data and 0.588 for RACE-H.

Methodology

Fine-Tuning

Used low rank adaptation (LoRA) to train 1.3% of the total 1.5B parameters (attention dimension = 16; α = 16, dropout = 0.05). Highlighting (HL) and Self-Assessment (SA) methods from [3] used for data augmentation. The best model and hyperparameter values from an augmented data grid search were then used to further train on the original, unaltered training and validation sets (learning rate = 1e-3, epochs = 5).

Experiments

1. Information Removal: Removed reading passage from the prompt to assess models' ability to learn patterns within question/answer choices alone.
2. Information Alteration: Randomly replaced prompt words with a random NLTK wordnet synset synonym to assess reliance on common vocabulary. Appended a random sentence from the previous passage to the end of the current passage to assess robustness to irrelevant information.
3. Chain-of-Thought: One- and two-shot prompting used to elicit model rationales. Responses evaluated with BERTScore, ROUGE, and human coherence ratings.
4. Rationality Profiling: Used a suite of interpretability metrics including Reasoning Score (robustness to token-level changes), Context Fidelity (prediction shifts under passage removal), and SHAP (token-wise attribution strength), to analyze model behavior beyond accuracy.

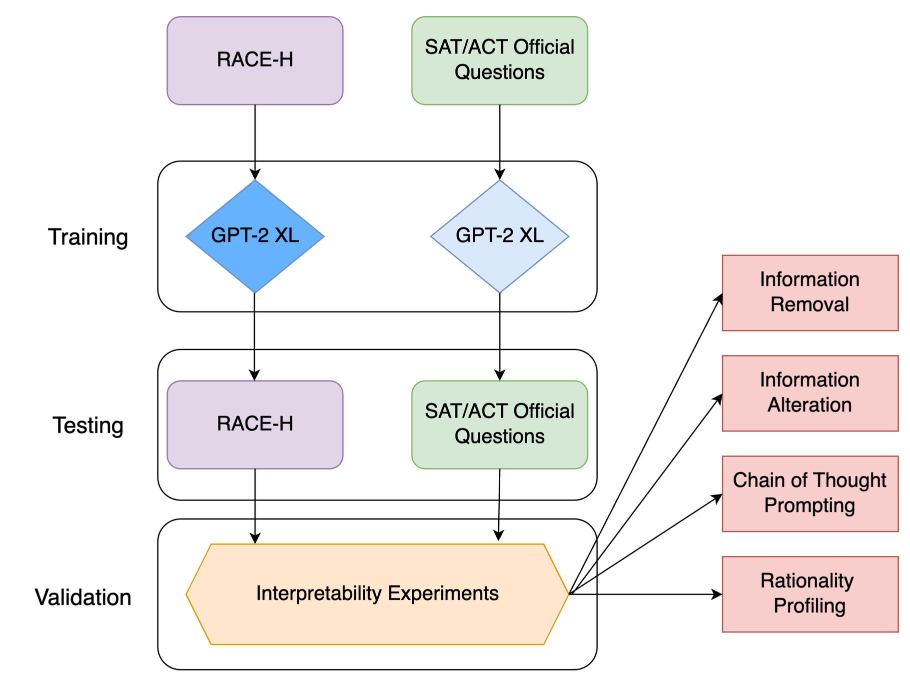

Figure 3. Overview of the experimental pipeline outlining the fine-tuning process, testing datasets, and interpretability experiments.

Results

Model	Dataset	Baseline	IR	IA - Syns	IA - Sent	Chain-of-Thought		Rationality Profiling
						BERTScoreF1	ROUGE-1	SHAP
Base	SAT/ACT	0.307	0.246	0.239	0.303	0.57	0.21	2.21
	RACE-H	0.312	0.241	0.246	0.307	0.55	0.32	1.67
SAT/ACT fine-tuned	SAT/ACT	0.622	0.292	0.251	0.618	0.52	0.27	2.99
RACE-H fine-tuned	RACE-H	0.420	0.260	0.246	0.402	0.51	0.22	2.52

Table 2. IR: Information Removal; IA - Syns: Information Alteration - Synonyms; IA - Sent: Information Alteration - Sentences

Model	Dataset	Average Human Evaluations
SAT/ACT FT	SAT/ACT	2.45
	RACE-H	2.41
RACE-H FT	SAT/ACT	1
	RACE-H	0.66

Table 3. CoT Human evaluation scores in the range of [0, 5]

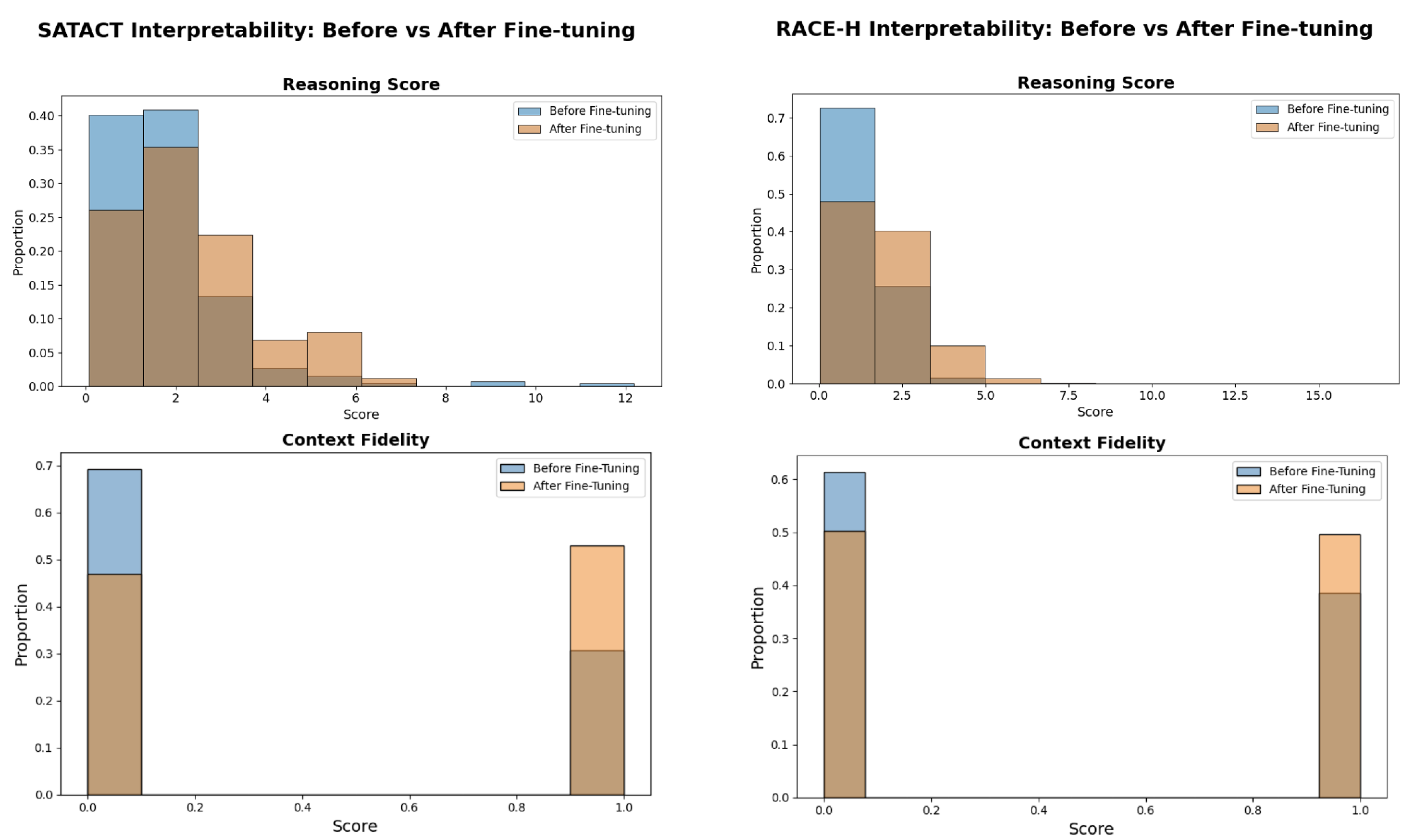

Figure 4. Interpretability metrics before and after fine-tuning on SAT/ACT and RACE-H datasets.

Discussion

SAT/ACT fine-tuned model:

- Improved accuracy and CoT rationale generation over base model
- Improved performance on IR and IA tasks – particularly IR
- Increased sensitivity to input features, as shown by SHAP analysis and token-level attribution shifts.

Take-Away Finding: Learning of test-specific features and increased reliance on exam vocabulary.

RACE-H fine-tuned model:

- Weaker gains post-fine-tuning
- Similar performance trends to SAT/ACT model

Take-Away Finding: Performance differences between the fine-tuned models despite comparable base performance indicate different strategies.

Chain-of-Thought Output

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.. . . . [

Future Work

- Scaling up trainable layers, LoRA attention dimension, and training data for performance gains
- Exploring other data augmentation strategies
- Experimenting with different strategies for defining model predictions for multiple-choice tasks
- Shifting learning objective to improve semantic similarity between CoT model generations and ground truth rationales
- Benchmarking the SAT/ACT dataset across other LLMs

References

[1] CollegeBoard.
Sat suite of assessments, 2025.
Official website of the SAT Suite of Assessments.

[2] ACT Inc.
Act - solutions for college and career readiness, 2025.
Official website of the ACT assessment.

[3] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie.
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, June 2019.