Answer-Level Calibration for SAT/ACT Multiple Choice Questions

Sandy Berrocal
Center for Data Science
New York University
sbb9435@nyu.edu

Anusha Dasgupta
Center for Data Science
New York University
ad7038@nyu.edu

Joy Fan
Center for Data Science
New York University
jf3511@nyu.edu

Sophie Juco
Center for Data Science
New York University
smj490@nyu.edu

Abstract

Standardized tests, such as the SAT and ACT, are critical components of college admissions in the United States, requiring students to develop strategies for answering complex multiple-choice questions across subjects like reading, writing, history, and science. This project evaluates the performance of a GPT-2 model on SAT and ACT multiple-choice questions and identifies key phrases in prompts and answers that lead the model to the correct answer. Approximately 1000 SAT and ACT practice questions were manually formatted into a compatible dataset from official CollegeBoard and ACT resources. We successfully implemented code to extract the most influential terms and phrases from the context, questions, and answer options. This analysis highlights the model's answering strategy and its potential application in assisting students to better understand effective approaches for tackling these standardized tests.

1 Introduction

The SAT and ACT are standardized tests widely used in US college admissions. These exams include context-based multiple-choice questions in subjects such as reading, writing, history, and science.

This project aims to implement a GPT-2 model for answering SAT/ACT-style multiple choice questions by applying answer-level calibration (ALC) techniques. While previous works have assessed NLP models for SAT/ACT math questions, no major attempts have addressed other subjects like English, History, and Science. Furthermore, there are no public datasets for this specific task, making this a novel application with a novel dataset.

The ALC model first evaluates answer choices without context to identify biases and their likelihood of being selected regardless of the question. It then evaluates the choices with the full question to predict which answer best fits the context, reducing biases like answer length and focusing on the content.

In addition to the ALC technique, the project implements a token saliency score extractor to identify key tokens in prompts and answer options that lead to correct predictions. These insights can aid question design and enhance students' understanding of standardized test strategies.

2 Related Work

This project is inspired by the ACL Anthology paper *Answer-level Calibration for Free-form Multiple Choice Question Answering* by Sawan Kumar [3]. Previous research has focused on leveraging pre-trained models (like BERT and RoBERTa) to improve question answering

performance through fine-tuning and techniques such as length normalization and probability calibration. However, these methods often fail to address biases from context-independent cues. Kumar's work incorporates prompting, calibration, and answer-level techniques to improve predictions for commonsense reasoning and dialogue-based questions.

While SAT questions have occasionally been used in related research, primarily for math or sentence completion tasks, we found no prior work applying ALC to SAT/ACT-style multiple-choice questions in the manner detailed above.

3 Approach

The existing model from Kumar's paper is a GPT-2 model. From the datasets used in Kumar's paper, we selected the SocialIQA dataset for pre-training since its question format is the same as the SAT/ACT data - includes context, question, multiple choice options, and correct answer label.

The SAT/ACT dataset was built by formatting practice questions from official websites into JSONL files with context, question, options, and the correct answer. Questions requiring table or graph interpretations were excluded as they would require an machine image processing and often were math-based questions. Additionally, underlined text in prompts were reformatted into explicit mentions, such as replacing "the underlined word" with "[specific word]" since the JSONL format does not encode text formatting. Since SocialIQA includes only three answer options per question, the model's code was adjusted to handle four options typical of SAT/ACT questions.

In addition to adapting the model for SAT/ACT questions, we built a component into the run code to extract the token-level saliency scores from the context, question, and options to assess which tokens contributed the most to the model's predictions. This analysis highlights key factors that drive correct answers, providing insights into the model's decision-making process.

4 Experiments

4.1 Data

We use SAT/ACT datasets with questions from subjects such as Science, English, and History, consisting of free-form multiple-choice questions with textual options of varying lengths.

Initially, we considered web/PDF scraping to collect SAT/ACT questions but found it unreliable. Instead, we manually built a JSONL dataset containing context, question, answer options, and the correct answer. The questions were sourced from CollegeBoard SAT Questions [2] and ACT Questions [1], resulting in approximately 1000 usable questions.

We also considered generating questions using tools like ChatGPT but opted against it due to concerns about the accuracy and reliability of generated questions.

4.2 Experimental details

For the trial of the training socialiqa data, we ran the model on one g2-standard-12 GPU which took about 30-40 minutes to complete. Besides changing the cache to our directories, we have kept the model and training configurations the same as they are in the existing model. The specific settings of this model are the use of a Causal Language Model (CLM), rather than a Masked Language Model (MLM), the use of gpt2-xl model, four processes used, and 100 episodes.

4.3 Discussion of Results

The tables compare the performance of different calibration methods applied to SAT/ACT and SocialIQA questions using accuracy and Expected Calibration Error (ECE) as metrics.

Table 1: Results Based on SAT/ACT Questions

Method	Train	ECE
answer_only	26.00	0.58
answer_only_worst	28.09	0.57
answer_only_norm	24.24	0.25
answer_only_worst_norm	26.81	0.23
uncalibrated	31.30	0.52
length_normalized	26.65	0.21
token_calibration	25.68	0.28
alc_unscaled	33.87	0.39
alc_tvd	32.10	0.42
alc_bc	31.14	0.41

Table 2: Results Based on SocialIQA Questions

Method	Dev	ECE
answer_only	36.18	0.52
answer_only_worst	32.45	0.58
answer_only_norm	38.43	0.31
answer_only_worst_norm	30.25	0.40
uncalibrated	40.53	0.47
length_normalized	41.35	0.28
token_calibration	34.65	0.45
alc_unscaled	42.63	0.36
alc_tvd	43.91	0.35
alc_bc	45.14	0.33

Calibration Methods:

- **Answer Only:** This baseline method evaluates answer choices without the context of the question, exposing biases like answer length or token frequency.
- **Answer Only Worst:** This method focuses on the worst-performing cases among answer-only predictions, identifies cases where bias has the biggest impact.
- **Answer Only Norm:** This method adjusts answer scores to reduce the influence of biases like answer length or word frequency, making the model's confidence more consistent.
- **Answer Only Worst Norm:** This method combines worst-case performance with normalization.
- **Uncalibrated:** This method evaluates the full input (context, question, and answer choices) without applying calibration adjustments, representing baseline performance for context-aware predictions.
- **Length Normalized:** This method adjusts predictions to address answer length bias.
- **Token Calibration:** This method corrects for token frequency bias.
- **ALC Unscaled:** This method adjusts the model's confidence scores to make predictions more reliable by reducing bias, without applying any scaling factors.
- **ALC TVD:** This method uses a mathematical way to measure the gap between confidence and correctness to make the confidence scores more accurate.
- **ALC BC:** This method applies Bayesian calibration, uses a method based on probability and prior knowledge to fine-tune the confidence scores for better accuracy.

Insights:

- **Accuracy:** For SAT/ACT questions, ALC methods, especially ALC Unscaled, perform best in accuracy, showing they effectively use context and reduce biases. For SocialIQA, ALC BC has the highest accuracy, demonstrating how Bayesian calibration helps handle more complex datasets.
- **Calibration Improvements:** Methods like Length Normalized and Token Calibration significantly reduce ECE, showing they are effective at fixing specific biases.
- **Trade-offs:** Length Normalized and Token Calibration reduce ECE but often lower accuracy. This trade-off is useful in situations where having more reliable confidence scores is more important than achieving the highest accuracy.
- **Dataset-Specific Observations:** SAT/ACT questions, with their complex contexts, benefit more from ALC methods compared to SocialIQA questions.

In conclusion, ALC methods provide a good balance by improving both accuracy and confidence reliability.

4.4 Interpretability Result

The model provides an output JSON file containing the context, question, answer choices, predicted answer, and their saliency scores. The JSON file is processed to extract and map the raw saliency scores to their associated tokens. The saliency score represents the importance of the token in the model's decision. In the formatted output, each token is presented once with the sum of its saliency scores.

4.4.1 Correct Example

An example of the saliency scores for a correct answer is presented in Appendix A. In this example, the model correctly predicts the option C, "to highlight two approaches to achieving political representation for Indigenous people," as the best match for the question. The predicted choice score (50.46) indicates a strong confidence level.

Overall, token scores are higher in the answer choices than in the prompt and question, indicating that the model relies more on the answer tokens than the context/question tokens. In the prompt and question, all of the question tokens are highly scored, while only keywords have high scores in the prompt. This indicates that the model highly considers what the question is asking. In particular, for the prompt, key tokens are ones associated with places or groups of people, such as "Indigenous" (0.181), "politicians" (0.164), "Ecuador" (0.224), "Latin" (0.149), "American" (0.102), and "Assembly" (0.216). "Indigenous" and "politicians" underline the main subject of the text, while "Latin America," "Ecuador," and "Assembly" highlight a specific example of Indigenous political success, which supports the theme of the text.

The saliency scores for the tokens in the answer choices reveal the reasoning behind the selection, as the correct choice has high scores for tokens like "highlight" (0.869) and "approaches" (0.704), which emphasize the comparison of different approaches. In comparison, the other answer choices lack this alignment. For example, Choice A: "trace the history," scores lower on "history" (0.327), as the context does not focus on historical development.

In general, the model ignores irrelevant words like "to" or "the" and focuses on action verbs and proper nouns connected to identifying the right answer. This strategy is similar to what is taught when preparing for standardized tests—ignoring unnecessary details in prompts and answer choices to focus on the most relevant information.

4.4.2 Incorrect Example

An example of the saliency scores for an incorrect answer is presented in Appendix B. In this example, the model incorrectly predicts the option B, "the human genome contains multiple transposons from the line family that are all primarily active in the hippocampus.," as the best match for the question rather than the correct option, A, "the line transposon in o. vulgaris and o. bimaculoides genomes is active in an octopus brain structure that functions similarly to the human hippocampus." The prediction score for choice B is significantly high at 90.77, suggesting strong model confidence in its incorrect answer. The overall strategy of focusing on proper nouns and action verbs seen in the correct example are seen in this example as well. For the context scores, tokens like "LINE" (0.601), "genomes" (0.231), and "hypothesis" (0.962) have high saliency scores, which indicates that the model recognizes the relevance of these terms to the question. However, the model fails to correctly relate the importance of the context describing "brain" (0.063) activity and its similarity to the human hippocampus, which is central to choice A. The model instead emphasizes "family" (0.196), which seems to align more with choice B ("transposons from the line family") rather than choice A. For choice scores, tokens in choice A, such as "brain" (0.148) structure (0.059)" and "functions (0.147) similarly (0.021) to the human hippocampus", are relevant to the hypothesis discussed in the context. However, the model does not appropriately prioritize these tokens, likely leading to the misprediction. The model might be overgeneralizing based on its pre-trained biases or prior examples where the "genomes" and "human hippocampus" association was more frequent. It fails to interpret that the active LINE transposons in octopus brain structures functioning like the human hippocampus directly support the hypothesis. In general, this indicates that the model lacks the semantic understanding needed to differentiate between options A and B which replicates the frequent human error of identifying two similarly correct answers and selecting the one that is not the "best" answer.

5 GitHub Repository

The GitHub Repository for this project can be found at the following link:
`https://github.com/sophiejuco/DS_GA_1011-Final_Project`

References

[1] ACT, Inc. Science practice test questions, 2024. Accessed: 2024-11-05.

[2] College Board. Sat suite question bank, 2024. Accessed: 2024-11-05.

[3] Sawan Kumar. Answer-level calibration for free-form multiple choice question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 665–679, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A Example of Saliency Scores for Correct Prediction

A.1 Context and Question

A number of Indigenous politicians have been elected to the United States Congress since 2000 as members of the country's two established political parties. In Canada and several Latin American countries, on the other hand, Indigenous people have formed their own political parties to advance candidates who will advocate for the interests of their communities. This movement has been particularly successful in Ecuador, where Guadalupe Llori, a member of the Indigenous party known as Pachakutik, was elected president of the National Assembly in 2021.

Question: Which choice best states the main purpose of the text?

A.2 Choices

1. to trace the history of an indigenous political movement and speculate about its future development
2. to argue that indigenous politicians in the united states should form their own political party
3. to highlight two approaches to achieving political representation for indigenous people
4. to consider how indigenous politicians in the united states have influenced indigenous politicians in canada and latin america

A.3 Predicted Answer

Predicted Choice Index: 2

Predicted Choice: to highlight two approaches to achieving political representation for indigenous people

Predicted Choice Score: 50.45880126953125

A.4 Context Saliency Scores (Token: Average Score)

```
A: 0.12866896204650402
number: 0.07837824150919914
of: 0.06964219082146883
Indigenous: 0.18118196725845337
politicians: 0.16389880515635014
have: 0.026370616164058447
been: 0.04801136255264282
elected: 0.1310657076537609
to: 0.015515659470111132
the: 0.15782203525304794
```

United: 0.05952558759599924
States: 0.0725106056779623
Congress: 0.09559346176683903
since: 0.060240404680371284
2000: 0.08285829424858093
as: 0.026250062976032495
members: 0.09271673858165741
country: 0.05238133855164051
âĢ: 0.16116104274988174
Ļ: 0.08622078038752079
s: 0.0330743882805109
two: 0.032407973892986774
established: 0.05993690621107817
political: 0.07029809802770615
parties: 0.061374871991574764
.: 0.11475667357444763
In: 0.04039040859788656
Canada: 0.10035440139472485
and: 0.021411452908068895
several: 0.0704238060861826
Latin: 0.1488421279937029
American: 0.10247173346579075
countries: 0.07203075103461742
,: 0.040723408572375774
on: 0.022723783738911152
other: 0.027907345443964005
hand: 0.05093321204185486
people: 0.07300614472478628
formed: 0.08719472773373127
their: 0.03741537220776081
own: 0.06557072978466749
advance: 0.06990860961377621
candidates: 0.06956566497683525
who: 0.04487685766071081
will: 0.050055207684636116
advocate: 0.08791783079504967
for: 0.016641330439597368
interests: 0.07354738190770149
communities: 0.06450211443006992
This: 0.05650158133357763
movement: 0.09638545103371143
has: 0.057469057850539684
particularly: 0.10685937665402889
successful: 0.07510950788855553
in: 0.03792643267661333
Ecuador: 0.2242361679673195
where: 0.04318720381706953
Gu: 0.05929513834416866
adal: 0.09972630068659782
upe: 0.09960883669555187
L: 0.03321884199976921
lor: 0.05693741422146559
i: 0.03631751053035259
a: 0.06047441530972719
member: 0.18493985012173653
party: 0.09128515236079693
known: 0.10276315361261368
P: 0.04099621903151274
ach: 0.041731405071914196

ak: 0.038980113342404366
ut: 0.04085979703813791
ik: 0.03634589817374945
was: 0.11396918632090092
president: 0.15528923645615578
National: 0.1331179030239582
Assembly: 0.21596834808588028
2021: 0.26878657564520836
Which: 0.275352343916893
choice: 0.19002841040492058
best: 0.17228573560714722
states: 0.2162502072751522
main: 0.21190103143453598
purpose: 0.16142840683460236
text: 1.0
is: 0.22580678015947342

A.5 Choice Saliency Scores (Token: Score)

: 0.7410133481025696
to: 0.18586720526218414
trace: 0.49504122138023376
the: 0.14521969854831696
history: 0.32687023282051086
of: 0.11589153856039047
an: 0.22384436428546906
indigenous: 0.624567985534668
political: 0.37551429867744446
movement: 0.27383744716644287
and: 0.07290558516979218
speculate: 0.7976685166358948
about: 0.13806205987930298
its: 0.08655235916376114
future: 0.3073637783527374
development: 0.0
argue: 0.33928540349006653
that: 0.14740394055843353
politicians: 0.1390933096408844
in: 0.08530501276254654
united: 0.7989344596862793
states: 0.7801619172096252
should: 0.24144864082336426
form: 0.09156119078397751
their: 0.03147493675351143
own: 0.08519447594881058
party: 0.0
highlight: 0.8688997626304626
two: 0.4092724323272705
approaches: 0.7043868899345398
achieving: 0.5909357070922852
representation: 0.46744346618652344
for: 0.06353087723255157
people: 0.0
consider: 0.4440244436264038
how: 0.2692895531654358
have: 0.13633102178573608
influenced: 0.9186235666275024
can: 0.16948500275611877
ada: 0.45422494411468506

```
lat: 0.07710563391447067
in: 0.019692324101924896
americ: 0.15547798573970795
a: 0.0
```

B Example of Saliency Scores for Incorrect Prediction

B.1 Context and Question

Although many transposons, DNA sequences that move within an organism's genome through shuffling or duplication, have become corrupted and inactive over time, those from the long interspersed nuclear elements (LINE) family appear to remain active in the genomes of some species. In humans, they are functionally important within the hippocampus, a brain structure that supports complex cognitive processes. When the results of molecular analysis of two species of octopus—an animal known for its intelligence—were announced in 2022, the confirmation of a LINE transposon in Octopus vulgaris and Octopus bimaculoides genomes prompted researchers to hypothesize that that transposon family is tied to a species' capacity for advanced cognition.

Question: Which finding, if true, would most directly support the researchers' hypothesis?

B.2 Choices

1. The LINE transposon in O. vulgaris and O. bimaculoides genomes is active in an octopus brain structure that functions similarly to the human hippocampus.
2. The human genome contains multiple transposons from the LINE family that are all primarily active in the hippocampus.
3. A consistent number of copies of LINE transposons is present across the genomes of most octopus species, with few known corruptions.
4. O. vulgaris and O. bimaculoides have smaller brains than humans do relative to body size, but their genomes contain sequences from a wider variety of transposon families.

B.3 Predicted Answer

Predicted Choice Index: 1

Predicted Choice: The human genome contains multiple transposons from the LINE family that are all primarily active in the hippocampus.

Predicted Choice Score: 90.76872253417969

B.4 Context Saliency Scores (Token: Average Score)

```
Although: 0.3205289766192436
many: 0.10691570304334164
trans: 0.15499895252287388
pos: 0.07360068894922733
ons: 0.11413001269102097
,: 0.1495254933834076
DNA: 0.1413827296346426
sequences: 0.12163253873586655
that: 0.09736684337258339
move: 0.0963746216148138
within: 0.054268548265099525
an: 0.08575733564794064
organism: 0.1320797074586153
âĢ: 0.5346675962209702
Ļ: 0.40683770179748535
s: 0.07564838789403439
genome: 0.1642252467572689
```

through: 0.06696477718651295
shuff: 0.11972715705633163
ling: 0.043421026319265366
or: 0.05811785068362951
duplication: 0.09970532357692719
have: 0.05272660031914711
become: 0.059485903941094875
corrupted: 0.15913963317871094
and: 0.055255225859582424
inactive: 0.108577661216259
over: 0.04698522202670574
time: 0.041532641276717186
those: 0.07386011816561222
from: 0.06389114912599325
the: 0.2219538912177086
long: 0.057102411054074764
inter: 0.08774622809141874
sp: 0.045964802615344524
ersed: 0.13572897389531136
nuclear: 0.12228952720761299
elements: 0.1263014329597354
(: 0.09177928045392036
LINE: 0.600860595703125
): 0.06942567694932222
family: 0.19630122557282448
appear: 0.06242042500525713
to: 0.061933585442602634
remain: 0.054140856489539146
active: 0.06952614337205887
in: 0.052540095522999763
genomes: 0.23075327649712563
of: 0.04944988805800676
some: 0.05836081504821777
species: 0.14403695613145828
.: 0.25655463337898254
In: 0.06101597938686609
humans: 0.14739885926246643
they: 0.04182522650808096
are: 0.03785637579858303
functionally: 0.10082598403096199
important: 0.06867998745292425
hippocampus: 0.2872670404613018
a: 0.08135777898132801
brain: 0.06309944204986095
structure: 0.07257370185106993
supports: 0.07681773602962494
complex: 0.04486938938498497
cognitive: 0.08488250337541103
processes: 0.05824098363518715
When: 0.09308844245970249
results: 0.06493582762777805
molecular: 0.08069431781768799
analysis: 0.07305681984871626
two: 0.05213169567286968
oct: 0.4000604096800089
opus: 0.09908062219619751
âĢĶ: 0.11127842403948307
an: 0.035121993627399206
animal: 0.07224258594214916

known: 0.06575390230864286
for: 0.09153653495013714
its: 0.04511718265712261
intelligence: 0.08117640763521194
were: 0.09081836976110935
announced: 0.12399527803063393
2022: 0.22899344190955162
confirmation: 0.08809598535299301
LINE: 0.4344278462231159
on: 0.06339630763977766
Oct: 0.11904050782322884
vulgar: 0.21785831451416016
is: 0.6104040890932083
b: 0.047235541976988316
im: 0.035043453332036734
ac: 0.03357093874365091
ulo: 0.10904175601899624
ides: 0.050230211578309536
prompted: 0.17173225060105324
researchers: 0.32152774930000305
hypothes: 0.33550963550806046
ize: 0.07939862459897995
is: 0.08323496021330357
tied: 0.16641755774617195
capacity: 0.17678572610020638
advanced: 0.14631101489067078
cognition: 0.2563389055430889
Which: 0.5862549766898155
finding: 0.32667670026421547
if: 0.24477654322981834
true: 0.26389430090785027
would: 0.3059023804962635
most: 0.15776906162500381
directly: 0.20685221627354622
support: 0.2810887545347214
hypothesis: 0.9616651386022568

B.5 Choice Saliency Scores (Token: Score)

: 0.7966068983078003
the: 0.043296195566654205
line: 1.0
trans: 0.09501440078020096
pos: 0.19386601448059082
on: 0.020482121035456657
in: 0.02505323849618435
o: 0.13293667137622833
.: 0.0
vulgar: 0.3055790960788727
is: 0.08949615806341171
and: 0.15540242195129395
b: 0.0771539956331253
im: 0.04947230592370033
ac: 0.0355973094701767
ulo: 0.16585607826709747
ides: 0.12860488891601562
genomes: 0.35904744267463684
is: 0.14807017147541046
active: 0.0410366952419281

an: 0.06401658803224564
oct: 0.3188284635543823
opus: 0.06846705824136734
brain: 0.14773958921432495
structure: 0.059177204966545105
that: 0.028282593935728073
functions: 0.14708736538887024
similarly: 0.0210320632904768
to: 0.08804161846637726
human: 0.3078693151473999
hippocampus: 0.029606951400637627
genome: 0.5881386995315552
contains: 0.18039393424987793
multiple: 0.17941580712795258
ons: 0.09618847817182541
from: 0.07012848556041718
family: 0.07798467576503754
are: 0.04663616046309471
all: 0.0735788345336914
primarily: 0.09845490008592606
a: 0.0806657001376152
consistent: 0.8644396662712097
number: 0.45028963685035706
of: 0.048354264348745346
copies: 0.42517974972724915
present: 0.1576707363128662
across: 0.19561854004859924
most: 0.16475358605384827
species: 0.07504940778017044
,: 0.033052168786525726
with: 0.06905671209096909
few: 0.15610750019550323
known: 0.14478737115859985
corrupt: 0.3999016284942627
ions: 0.016657663509249687
have: 0.48910680413246155
smaller: 0.3473582863807678
brains: 0.29452940821647644
than: 0.1055246889591217
humans: 0.1909603327512741
do: 0.0792231485247612
relative: 0.1811894029378891
body: 0.07337728142738342
size: 0.04843689873814583
but: 0.06602087616920471
their: 0.0728231817483902
contain: 0.10339801758527756
sequences: 0.2615143358707428
wider: 0.17474451661109924
variety: 0.051804181188344955
families: 0.03439094498753548