Answer-level Calibration for Free-form Multiple Choice Question Answering

Sawan Kumar
Indian Institute of Science, Bangalore
sawankumar@iisc.ac.in

Abstract

Pre-trained language models have recently shown that training on large corpora using the language modeling objective enables few-shot and zero-shot capabilities on a variety of NLP tasks, including commonsense reasoning tasks. This is achieved using text interactions with the model, usually by posing the task as a natural language text completion problem. While using language model probabilities to obtain task specific scores has been generally useful, it often requires task-specific heuristics such as length normalization, or probability calibration. In this work, we consider the question answering format, where we need to choose from a set of (free-form) textual choices of unspecified lengths given a context. We present ALC (Answer-Level Calibration), where our main suggestion is to model context-independent biases in terms of the probability of a choice without the associated context and to subsequently remove it using an unsupervised estimate of similarity with the full context. We show that our unsupervised answer-level calibration consistently improves over or is competitive with baselines using standard evaluation metrics on a variety of tasks including commonsense reasoning tasks. Further, we show that popular datasets potentially favor models biased towards easy cues which are available independent of the context. We analyze such biases using an associated F1-score. Our analysis indicates that answer-level calibration is able to remove such biases and leads to a more robust measure of model capability.

1 Introduction

Language models (LM), trained on large corpora, have been shown to exhibit few-shot and zero-shot learning capability (Radford et al., 2019; Brown et al., 2020) using only text interactions, as opposed to finetuning the model parameters using task specific training examples. Relying purely on text interactions for few-shot ability shifts the focus to designing and utilizing suitable task-specific natural language templates.

In this work, we focus on free-form multiple choice question answering (and commonsense reasoning tasks in particular), where given a context and a set of choices of unspecified lengths, a model is required to select the most suitable choice. To enable zero-shot learning, the typical approach is to form textual sequences by concatenating the context independently with each choice and then scoring the concatenated strings using a pre-trained LM.

While LM probabilities have been shown to provide useful estimates of choice probabilities given a context, there is no incentive to treat the choices as equal in the absence of the associated context. For example, the LM probabilities in a neutral context are likely to be determined by frequency. In this work, we explore the role of biases that are likely to be associated with the choices naturally due to the language modeling objective. We propose ALC[1] (Answer-Level Calibration), where we use a neutral context to model such biases and remove them using a scaling factor determined by how similarly a model handles the question context as compared to a neutral context.

Further, we show that popular datasets favor models which rely on easy cues which are context independent. We use a bias-specific F1 score to analyze such biases. Our results indicate the need for answer-level calibration for more accurate estimates of model capabilities, or equivalently the design of better datasets. We hope our work will be useful for further research in both those directions. Specifically, we analyze context-independent biases related to length, part-of-speech (POS) and neutral context probabilities of the choices.

In summary, we make the following contributions:

[1] ALC source code is available at https://github.com/SawanKumar28/alc

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 665 - 679
May 22-27, 2022 ©2022 Association for Computational Linguistics

1. We present ALC, a model-agnostic approach to improve the unsupervised performance of pretrained LMs for free-form multiple choice question answering, including commonsense reasoning tasks.
2. We show that popular datasets favor models relying on context-independent easy cues and demonstrate the need for answer-level calibration to better estimate model capabilities.

2 Related Work

Prompts Jiang et al. (2020) show that manually created templates can be sub-optimal in extracting knowledge from LMs, and propose mining and paraphrasing-based approaches using training examples. Schick and Schütze (2021) highlight the importance of selecting templates for enabling few-shot learning.

Calibration Probabilities output by neural networks are known to suffer from lack of calibration (Guo et al., 2017), including LM output probabilities (Braverman et al., 2020). Zhao et al. (2021) use token-level calibration to improve on few-shot classification and generation tasks. In contrast, we show that answer-level calibration is more suitable for the multiple choice setting that we consider.

While we focus on free-form multiple choice questions in this work, when the choices are single tokens, for example in a classification task where the choices are *True* and *False*, answer-level calibration would behave similar to token-level calibration. As a result, answer-level calibration can be seen to have a more general scope as also illustrated empirically through our experiments.

Further, our analysis (Section 3.4) shows that answer-level calibration provides a more reliable measure of model performance on datasets with potential biases.

Finally, Jiang et al. (2021) explore supervised methods, including finetuning as well as post-hoc methods, to improve calibration using training examples. In this work, we focus mainly on unsupervised calibration.

Answer-level calibration Brown et al. (2020) generally perform length normalization over the token probabilities for a choice, while observing that for a select few tasks they obtain performance gains when using an answer-level calibration scheme (which corresponds to the unscaled version in Equation 3 of ALC). They use task specific development sets to choose between length normalization and answer-level calibration which is undesirable for few-shot learning (Kann et al., 2019), and specifically for zero-shot learning. In this work, we show that unscaled calibration (as in Equation 3) is sub-optimal, compared to our proposed scaled version.

More recently, Holtzman et al. (2021) also arrive at a formulation equivalent to the unscaled version of ALC but are motivated differently. Specifically, they hypothesize that the possibility of different surface forms of the same concept causes a competition between surface forms when scored by the LM. In contrast, we are motivated by calibration concerns and the presence of context-independent biases. We justify this motivation through bias associated evaluation (Section 5.2) for both the unscaled and scaled versions of ALC.

Alternative approaches using enhanced context One way to make the probability estimates of the choices more accurate is to enhance the context using more task-specific cues. For example, Brown et al. (2020) show that with just a few in-context examples, significant gains in performance can be obtained. At the same time, it has been shown that the order of examples as well as token-level calibration in such prompts can be critical for getting good performance (Zhao et al., 2021; Kumar and Talukdar, 2021).

While the gains from enhanced context through additional examples may be complementary to answer-level calibration, we focus on the zero-shot setting in this work. In the zero-shot setting, Shwartz et al. (2020), working on the question answering format, propose generating textual clarifications using the pre-trained LM itself, to enhance the context and improve zero-shot performance of pre-trained LM on commonsense reasoning tasks. While their method has a much higher computational cost, we use it as an unsupervised baseline and show improvement over it on most tasks we consider.

3 ALC: Proposed Method

We introduce the problem setting and notation in Section 3.1. We briefly describe our motivation in Section 3.2 and discuss the core idea of removing context-independent biases in Section 3.3. We provide the natural language formatting used in our experiments in Section A.3. We discuss bias associated measures in Section 3.4.

3.1 Notation

We consider a problem setting where an example consists of a textual context C and K textual choices (or options) $O_k, k \in [K]$, and we need to predict which choice O_k fits best in context C. For example, in the case of question answering, this amounts to answering a question contained in the context C. Additionally, we define an instance-independent neutral context C_ϕ, where we expect all choices to be equally likely.

Denoting the gold answer by Y, the evaluation data is comprised of N instances defined by the tuples $(C^i, [O^i_k], Y^i), k \in [K], i \in [N]$.

3.2 Motivation

Our main motivation is to evaluate the suitability of pretrained LMs for free-form multiple choice question answering where we contend that raw conditional phrase probabilities do not satisfy a natural requirement for such tasks (Equation 2). We suggest and evaluate modifications to meet this requirement.

3.3 Removing Context-independent Biases

We aim to obtain a probabilistic model M which provides estimates $P_M(O|C)$, the probability of a choice O given the context C. Predictions y for an example can subsequently be made using:

$$y = \mathrm{argmax}_k(P_M(O_k|C)) \tag{1}$$

We wish to build such a model using a pretrained LM, e.g., GPT2. Such a LM, trained on the task of next word prediction, is expected to provide estimates of word probabilities given a textual context. For example, given the sequence of words $w_1 w_2 ... w_i$, we expect GPT2 to provide probability estimates $P_L(w_{i+1}|w_1 w_2 ... w_i)$. Applying chain rule, we can obtain estimates of phrases given a textual context. For example, we could obtain estimates of $P_L(O|C)$.

Can $P_L(O|C)$ serve as a proxy for $P_M(O|C)$? It is tempting to expect the LM probabilities $P_L(O|C)$ to serve as a proxy for $P_M(O|C)$ when we can format the task in natural language. However, under the assumption that all choices O_k are equally likely given a neutral context C_ϕ, this approximation can be sub-optimal. For it to be optimal, we would need

$$P_L(O_1|C_\phi) = P_L(O_2|C_\phi) ... = P_L(O_K|C_\phi) \tag{2}$$

However, given that these are task and instance specific choices, there is no incentive in the language modeling objective to ensure this condition.

To address this, we define a new score $S_L(O|C)$ to behave as expected with a neutral context:

$$S_L(O_k|C) = \log P_L(O_k|C) - \log P_L(O_k|C_\phi) \tag{3}$$

Predictions can subsequently be made using:

$$y' = \mathrm{argmax}_k(S_L(O_k|C)) \tag{4}$$

Scaling the bias term: Equation 3, while desirable, makes a strong assumption about how the bias is present in the LM. While valid unquestionably for the neutral context, the bias in a trained (on task-specific data, or on a task-independent pretraining corpus) model is likely to depend on the context as well. For instance, a longer or more familiar context (in terms of similarity to training contexts) may mean the model is less reliant on context-independent cues. We therefore define a scaled version for removing biases, where the function g outputs the scaling term (ranging in $[0, 1]$):

$$S'_L(O_k|C) = \\ \log P_L(O_k|C) - g(C, C_\phi) * \log P_L(O_k|C_\phi) \tag{5}$$

We would want this formulation to preserve the requirement in Equation 2 which was satisfied by the unscaled version in Equation 3. Specifically, we want $g(C_\phi, C_\phi) = 1$ which would assign an equal score to each choice O_k given a neutral context.

To get a model-agnostic[2] estimate of g, we think of $\log P_L(O_k|C)$ and $\log P_L(O_k|C_\phi)$ as outputs from different models M and M_ϕ respectively, and g as a measure of similarity between the models. Note that while M uses the available context C, M_ϕ uses only the neutral context C_ϕ. The intuition is that if M and M_ϕ are identical, there is no new information provided by M and we want to set $g(C, C_\phi) = 1$, leading to $S'_L(O_k|C) = 0$. On the other hand, if M and M_ϕ are very dissimilar, we can rely on the contextual scores of M and set $g(C, C_\phi) = 0$, leading to $S'_L(O_k|C) = \log P_L(O_k|C)$. Specifically, to estimate g, we compute a similarity metric between the token probabilities (across the model's entire vocabulary) output by the two models.

[2] By model-agnostic, we mean we only access the probabilities output by the model and don't rely on any knowledge of the model architecture.

$$g(C, C_\phi) = \text{Sim}(p_L^f(C), p_L^f(C_\phi)) \quad (6)$$

where p_L^f indicates the probability vector output by the model across the vocabulary for the first token given the corresponding context. In this work, we consider Total Variation Distance (TVD), and Bhattacharyya Coefficient (BC) (Bhattacharyya, 1943).

When using TVD, we subtract it from 1, to obtain a similarity estimate:

$$g^{\text{TVD}}(C, C_\phi) = 1 - 0.5 * ||p_L^f(C) - p_L^f(C_\phi)||_1 \quad (7)$$

while we directly use BC:

$$g^{\text{BC}}(C, C_\phi) = \sum_{i=1}^{V} \sqrt{p_L^f(C)[i] * p_L^f(C_\phi)[i]} \quad (8)$$

3.4 Bias Associated Measures

Consider an instance and choice specific attribute $A^i(O_k^i)$ which can take values $a_j, j \in [J]$. If we expect the attribute to be uncorrelated with task performance, we expect a model to perform similarly when evaluating subsets with different distributions of attributes $A^i(O_k) = a_j$. If a model relies on specific values of the attribute and if the evaluation data has sufficient representation of that value, standard evaluation metrics which ignore this attribute may provide an erroneous estimate of the model capability. As an extreme example, consider $A(.)$ to denote whether the selected choice corresponds to the shortest choice among all choice $O_k, k \in [K]$, with the attribute values being true/false. Assume then that the evaluation data is dominated by instances where $A^i(Y^i) = \text{true}$, i.e., with a high probability, the correct answer in the evaluation data is the shortest choice. Consider also a model which always chooses the shortest choice, irrespective of the content. The model would return close to perfect scores using standard evaluation metrics such as accuracy against gold labels.

To analyze the impact of such attributes, we use a macro F1 score which takes into account the partitions created by an attribute. Recalling that an instance is represented by the tuple $(C^i, [O_k^i], Y^i), i \in [N]$, and letting $\hat{Y}^i$ be the model prediction, we define precision (P), recall (R) and F1 scores for each attribute value a_j, and subsequently an attribute specific macro F1 score (F1_A).

$$P_{(A,a_j)} = \frac{\#\{(A^i(\hat{Y}^i) = a_j) \;\&\; (\hat{Y}^i = Y^i)\}}{\#\{A^i(\hat{Y}^i) = a_j\}} \quad (9)$$

$$R_{(A,a_j)} = \frac{\#\{(A^i(\hat{Y}^i) = a_j) \;\&\; (\hat{Y}^i = Y^i)\}}{\#\{A^i(Y^i) = a_j\}} \quad (10)$$

$$\text{F1}_{(A,a_j)} = \frac{2 * P_{(A,a_j)} * R_{(A,a_j)}}{P_{(A,a_j)} + R_{(A,a_j)}} \quad (11)$$

$$\text{F1}_A = \text{Average}(\{\text{F1}_{(A,a_j)}\}) \quad (12)$$

where $\#\{.\}$ denotes the count of the corresponding set. If the model performs similarly irrespective of the attribute value, the macro F1 score F1_A is equal to the standard measure of accuracy:

$$\text{Accuracy} = \frac{\#\{\hat{Y}^i = Y^i\}}{N} \quad (13)$$

4 Experimental Setup

The datasets used and the corresponding prompts are described in Section 4.1. The LMs used are described in Section 4.2 and the baseline approaches in Section 4.3. Experimental results and analyses are presented in Section 5.

4.1 Data

We used a series of commonsense reasoning tasks and evaluated on the publicly available development sets. We used the same versions of the data as Shwartz et al. (2020) to allow for a direct comparison — COPA (Gordon et al., 2012), CommonsenseQA (Talmor et al., 2019), MCTACO (Zhou et al., 2019), SocialIQA (Sap et al., 2019), PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2020). We also report on the adversarially generated large-scale SWAG dataset (Zellers et al., 2018).

Further, we report on the AI2 Reasoning Challenge (ARC) (Clark et al., 2018), which has Easy and Challenge versions.

As a representative dialog understanding task, we report on the DREAM (Sun et al., 2019) dataset.

Finally, we report on a recent benchmark introduced for measuring multitask accuracy of pretrained models (referred to as Hendrycks in the following) Hendrycks et al. (2020).

For MCTACO, we used a reduced subset as provided by Shwartz et al. (2020) where each question

Model		COPA	COPA-test	CSQA	MCTACO	MCTACO-test	SocialIQA	PIQA	WG
Accuracy with gpt2-xl (zero-shot)									
Majority		55.0	50.0	20.9	50.0	51.32	33.6	50.5	50.4
Self-talk		58.0	-	31.4	59.9	-	**46.2**	70.1	53.9
Token calibration		57.00	58.60	27.44	52.86	55.05	36.23	60.07	51.62
Uncalibrated		72.00	74.20	37.18	61.89	65.61	40.53	70.67	**55.49**
Length normalized		68.00	72.80	33.82	55.73	56.05	41.35	71.33	55.01
ALC	Unscaled	70.00	79.20	47.91	57.05	55.87	42.68	59.96	52.80
	TVD	**74.00**	**81.60**	46.19	64.54	**65.97**	43.91	**71.60**	55.25
	BC	73.00	80.00	**49.71**	**64.76**	64.60	45.14	70.78	54.06
Average gain in accuracy across LMs (zero-shot)									
Length normalized		-2.60	-0.68	-1.52	-4.05	-7.97	1.55	-0.50	-0.09
ALC	Unscaled	-4.20	5.04	8.14	-6.56	-10.03	2.82	-9.43	-0.35
	TVD	**2.80**	5.64	7.67	2.56	**1.64**	3.06	**0.12**	**0.43**
	BC	-1.00	**6.88**	**10.37**	**2.86**	-0.66	**4.08**	-0.87	0.38
Accuracy with gpt2-xl (1-shot)									
Length normalized		-	-1.54	0.91	-	-3.13	2.65	0.45	-0.50
ALC	Unscaled	-	3.55	6.96	-	-13.84	2.45	-9.31	-2.03
	TVD	-	4.58	6.13	-	**1.84**	3.62	**0.92**	-0.35
	BC	-	**5.43**	**9.72**	-	1.32	**3.77**	-0.07	-0.90
Accuracy with gpt2-xl (4-shot)									
Length normalized		-	-0.63	1.74	-	-0.79	4.46	0.43	-0.28
ALC	Unscaled	-	3.85	8.67	-	-11.30	3.22	-9.02	-2.05
	TVD	-	4.71	5.19	-	**2.43**	4.27	**1.00**	-0.63
	BC	-	**5.53**	**9.32**	-	2.25	**4.47**	0.32	-1.09

Table 1: *Standard evaluation results on unsupervised commonsense question answering tasks:* (**Top**) Dev set accuracies (unless specified otherwise) with gpt2-xl are presented for baselines and ALC along with an unscaled version of ALC where the bias term is not scaled. The highest accuracies are marked in bold font. Note that while the unscaled version provides gains over the uncalibrated baseline, on the CommonsenseQA and SocialIQA tasks, there is also a drop in performance on some datasets, notably on PIQA. The scaled version, on the other hand, outperforms the LM-Baseline on all datasets except WinoGrande (on which all models perform close to majority accuracy). While token calibration improves over the majority accuracy on all datasets, it performs worse than the uncalibrated baseline. Finally, ALC outperforms or is competitive with Self-talk, while being computationally more efficient. (**Middle**) We also report on the gain over the uncalibrated baseline over different gpt2 variants and observe similar trends. (**Bottom**) Finally, we report on few-shot evaluation with gpt2-xl and again observe similar trends. Please see Section 5.1 for more details.

is associated with only one correct choice. For COPA, we also report on the test split due to the small size of COPA dev set. The sizes of the datasets used are reported in Appendix Table 8. All datasets contain questions in English language. We briefly describe these datasets in Section A.2. Examples for each dataset along with contextual (C) and neutral (C_ϕ) prompts used in this work are captured in Section A.3.

4.2 Models

We experiment with GPT2 (Radford et al., 2019) variants - distilgpt2, gpt-small, gpt-medium, gpt2-large and gpt2-xl. The size of models used is reported in Appendix Table 9. While the gpt-* models have been trained similarly as described in Radford et al. (2019), distilgpt2 has been pretrained with the supervision of GPT2[3] (Wolf et al., 2020). For most of our experiments, we utilize the gpt2-xl model.

Please refer Section A.1 for additional details about the experimental setup.

4.3 Baselines

Uncalibrated: Predictions are made using uncalibrated probabilities from a LM, $\log P_L(O|C)$, computed as the sum of conditional log-probabilities output by the model for the tokens in O.

Length normalized: Predictions are made using length-normalized probabilities from a LM, $\log P_L(O|C)$, computed as the mean of conditional log-probabilities output by the model for the tokens in O.

[3] https://huggingface.co/distilgpt2

Model		ARC		DREAM	SWAG	Hendrycks-test			
		Easy	Challenge			Humanities	STEM	Social sciences	Other
Token calibration		35.09	20.40	40.20	29.53	23.38	22.70	25.45	25.51
Uncalibrated		58.25	27.76	48.14	49.30	26.99	24.16	31.52	31.55
Length normalized		50.70	29.43	48.77	**65.36**	29.33	26.47	30.84	32.85
ALC	Unscaled	53.33	33.11	52.99	57.04	**31.05**	**29.13**	**32.76**	**35.26**
	TVD	**60.00**	29.43	52.50	53.77	28.80	25.98	32.24	33.07
	BC	56.49	**33.78**	**53.14**	59.16	30.31	27.60	32.60	34.58

Table 2: *Standard evaluation results on additional tasks*: Dev set accuracies (unless specified otherwise) are reported. The trends are similar to Table 1, except for SWAG (see Section 5.2 and Table 6 for an explanation). Please see Section 5.1 for more details.

Model	Shortest=true			Longest=True		
	P	R	F1	P	R	F1
Commonsenseqa						
Size	213			266		
Uncalibrated	25.94	51.64	34.54	49.63	25.19	33.42
Length normalized	41.3	17.84	24.92	28.94	55.26	37.98
ALC (Unscaled)	48.91	31.46	38.29	46.2	57.14	51.09
ALC (BC)	42.27	38.5	40.29	51.16	49.62	50.38
SocialIQA						
Size	665			667		
Uncalibrated	38.27	68.72	49.17	48.36	15.21	23.15
Length normalized	51.82	10.68	17.71	38.69	78.29	51.78
ALC (Unscaled)	49.2	27.82	35.54	40.76	58.94	48.19
ALC (BC)	47.92	38.05	42.41	44.06	52.58	47.95

Table 3: *Length bias analysis:* We consider subsets of data where the shortest/longest choice is correct, and report on P, R and F1 scores (lowest values are underlined). We compare ALC against the uncalibrated as well as length normalized baselines. While length normalization is commonly used to overcome length bias, we find that it overcompensates and severely penalizes short answers (see recall with Shortest=true; the recall is lower than that for a random baseline). On the other hand, the uncalibrated baseline severely penalizes longer answers as expected. ALC improves on both subsets and provides a better alternative to length normalization. Please see Section 5.2.1 for details.

Self-talk: We use the official code repository[4] of self-talk (Shwartz et al., 2020) using gpt2-xl as both the scoring model and the knowledge source.

Token calibration: Following Zhao et al. (2021), we use the probability vector output, p_N^f by the model at the first token given the neutral context to calibrate the model probabilities. Specifically, each token probability p is offset by p_N^f and renormalized: $p' = \text{softmax}(p - p_N^f)$. We also tried an alternative variant suggested by Zhao et al. (2021) where $p' = \text{softmax}(p/p_N^f)$ but this generally did worse and we skip the corresponding results.

5 Experimental Results

We aim to answer the following questions:

Q1 How does ALC compare with baselines using standard evaluation (accuracy) on free-form multiple choice question answering tasks? (Section 5.1)

Q2 Does the aforementioned evaluation reflect true model capability? To answer this question, we perform a series of bias associated evaluations (see Section 3.4) and also evaluate whether ALC helps overcome such biases. Specifically, we evaluate on biases related to answer length, POS tag and context-ignorant LM probability. (Section 5.2)

Q3 Does ALC improve expected calibration error (Guo et al., 2017)? (Section 5.3)

5.1 Standard Evaluation

The overall results for the commonsense reasoning tasks (considered by Shwartz et al. (2020)) using standard evaluation of ALC, as well as the base-

[4] https://github.com/vered1986/self_talk

Model	POS = noun			POS = verb			POS = adj		
	P	R	F1	P	R	F1	P	R	F1
Commonsenseqa									
Size	902			149			142		
Uncalibrated	36.93	39.47	38.16	39.57	36.91	38.19	40.48	23.94	30.09
Length normalized	35.07	31.37	33.12	33	44.97	38.07	33.33	33.8	33.57
ALC (Unscaled)	48.68	45.01	46.77	43.75	56.38	49.27	52.74	54.23	53.47
ALC (BC)	49.32	48.34	48.82	48.84	56.38	52.34	59.32	49.3	53.85

Table 4: *POS bias analysis:* We consider subsets of data using the POS tag of the first token and report on P, R and F1 scores (lowest values are underlined). We limit to the larger subsets of nouns, verbs and adjectives (adj). We note that the uncalibrated baseline does worse on adjectives when compared to nouns and verbs. Both length normalized and ALC provide more even scores across POS tags. Please see Section 5.2.2 for details.

lines, with gpt2-xl are presented in Table 1 (top). We also report on an unscaled ablation of ALC. Note that ALC outperforms the uncalibrated baseline on all datasets except WinoGrande (where all models perform poorly and we drop it from further discussions). Further, the significant gains compared to token calibration (which generally does worse than the uncalibrated baseline) show that answer-level calibration is more suited for unsupervised commonsense question answering when there is no constraint on the lengths of candidate choices. Finally, ALC outperforms or is competitive with self-talk[5] while being significantly less computationally intensive. ALC requires scoring two strings (context input and neutral input) for each choice, while self-talk requires generating hundreds of clarification texts using data-dependent templates and subsequently scoring them.

We also report on the average gain over the uncalibrated baseline across gpt2 models of varying sizes (Table 9) in Table 1 (middle) and observe similar trends as in the case of gpt2-xl.

While our focus is zero-shot unsupervised evaluation, we also perform few-shot (1-shot and 4-shot) evaluation In general, for k-shot evaluation, we sample 100 sets of size k from an unseen split[6] of the dataset. A few-shot context is obtained by concatenating training examples with a newline token. We report the average performance on the evaluation set in Table 1 (bottom) and observe similar trends as before.

We present the standard zero-shot evaluation on additional datasets in Table 2. The trends are similar except for the SWAG (see Section 5.2 for an explanation) and the Hendrycks datasets (see Table 11).

Finally, while our focus is causal language models, we also present results using RoBERTa-large (a masked language model) in Table 10. Again, we observe similar trends.

In the subsequent sections, we show that the evaluation using the accuracy metric may not reveal true model capabilities as the datasets may favor models which utilize easy cues for predicting the answer.

5.2 Bias Associated Evaluation

Next, to gain a better understanding of the model capabilities, we analyze the performance associated with undesirable biases related to length, POS tag and context-ignorant LM probability. Specifically, we define the following attributes (see Section 3.4):

Shortest Attribute $A^i(O^i_k)$ is set to true if O^i_k is the shortest (number of tokens) choice among the choices $O^i_{k'}, k' \in [K]$. Otherwise, the attribute is set to false.

Longest Defined similar to Shortest, but set to true if O^i_k is the longest answer and false otherwise.

POS Attribute $A^i(O^i_k)$ is set to the POS tag of the first token in the choice O^i_k. We don't consider POS tags which occur less than a threshold (25) in the evaluation data.

LM-Best Attribute $A^i(O^i_k)$ is set to true if O^i_k is the most likely choice using context-ignore (neutral input) LM probability. Otherwise, it is set to false.

LM-Worst Defined similar to LM-Best, but set to true when O^i_k is the least likely choice and false otherwise.

Finally, we consider length-normalized versions of LM-Best and LM-Worst, referred to as **LM-**

[5] Please see Section A.4 for a note explaining the unusually high relative performance of baselines on some tasks when compared to self-talk.

[6] For few-shot evaluation, we sample from the training split for all except COPA and MCTACO datasets where we sample from the dev set and report on the test set.

Model	LM-Best = true			LM-Worst = True		
	P	R	F1	P	R	F1
PIQA						
Size	1195			643		
Uncalibrated	70.53	94.31	80.7	71.67	26.75	38.96
Length normalized	73.96	86.28	79.64	63.06	43.55	51.52
ALC (Unscaled)	77.42	54.23	63.78	45.35	70.61	55.23
ALC (BC)	75.74	81	78.29	59.46	51.79	55.36
ARC (Easy)						
Size	183			109		
Baseline	52.96	83.06	64.68	72.22	23.85	35.86
Length normalized	60.56	59.56	60.06	39.47	41.28	40.36
ALC (Unscaled)	81.71	36.61	50.57	36.2	73.39	48.48
ALC (BC)	64.15	55.74	59.65	45.04	54.13	49.17
ARC (Challenge)						
Size	64			86		
Uncalibrated	23.78	68.75	35.34	28.57	4.65	8
Length normalized	27.55	42.19	33.33	29.03	20.93	24.32
ALC (Unscaled)	38.71	18.75	25.26	33.87	48.84	40
ALC (BC)	29.76	39.06	33.78	36.99	31.4	33.96

Table 5: *Context-ignorant LM bias analysis:* We consider subsets of data where the correct choice corresponds to the best/worst choice as per the context-ignorant (neutral context) LM probability and report on P, R and F1 scores (lowest values are underlined). Note that the F1 performance of uncalibrated and length-normalized baselines on PIQA and ARC (Easy) is much higher when LM-Best=true, i.e., when the correct choice is also the most likely choice without considering the context. An important takeaway here is that while standard evaluation did not distinguish ALC from the baselines, ALC is not overly reliant on context-ignorant LM probabilities. Please see Section 5.2.3 for more details.

Norm-Best and **LM-Norm-Worst** respectively.

Briefly, our experiments reveal that while the datasets considered don't share a similar bias pattern, each usually suffers from at least one bias considered in this work, i.e., there is a drop in performance when measured using the bias associated score. We present the detailed results for commonsense reasoning tasks in Appendix Table 12, using gpt2-xl model, while highlighting the key takeaways here. Recall that in the absence of biases in the model, the F1 score should match the accuracy score.

In the following sections, we provide a more directed analysis on the presence of such biases, on datasets where such biases are most prominent, and if ALC helps alleviate such biases.

5.2.1 Length

We create subsets of the CommonsenseQA and SocialIQA dev set with specific properties to evaluate if the LM-Baseline has the associated biases and if they are addressed by ALC. First, we create subsets of examples where the shortest/longest answer is the correct answer. We expect longer sentences to have lower probabilities than shorter sentences with the uncalibrated baseline. Additionally, with the length normalized variant, where the final score is obtained as the mean of conditional log-probabilities instead of the sum (as in the uncalibrated baseline), longer sentences could potentially be favored. We report the uncalibrated and ALC's performance in Table 3. Note that both uncalibrated baseline and the length-normalized variants favor one subset at the cost of the other, while ALC improves on both. In particular, the uncalibrated baseline has a much poorer recall when the longest answer is correct. On the other hand, the length normalized variant has a much poorer recall when the shortest answer is correct. The results indicate that ALC provides a viable alternative to length normalization for handling length biases.

5.2.2 POS

We analyze potential part of speech (POS) tag biases in Table 4. Considering the CommonsenseQA dataset, we create subsets of the data where the correct answer is of the POS tag noun, verb or adjective. Note that ALC shows less variation in performance (F1) across these subsets when compared to uncalibrated baseline while improving on

Model	PIQA	SWAG
LM-Norm-Best		
Baseline	65.38	49.38
Length normalized	60.16	**60.97**
ALC (Unscaled)	58.89	58.12
ALC (BC)	**67.92**	59.48
LM-Norm-Worst		
Baseline	65.38	39.90
Length normalized	60.16	46.99
ALC (Unscaled)	58.89	49.48
ALC (BC)	**67.92**	**51.14**

Table 6: *Context-ignorant normalized scores:* LM-Norm-Best (top) and LM-Norm-Worst (bottom) macro F1 evaluation on SWAG and PIQA datasets (lowest values are underlined and highest values are in bold). The macro F1 scores for LM-Norm-Best and LM-Norm-Worst are identical for PIQA as the dataset contains only two candidate answers for a question and the subsets created by the two measures are identical. Note that while length-normalization has a higher accuracy than ALC on the SWAG dataset (Table 2), it does worse than ALC on the LM-Norm-Worst F1 score. Please see Section 5.2.3 for more details.

each subset. In particular, the maximum difference in F1 scores is 8.1 for the uncalibrated baseline while it is 5.03 for ALC (BC). ALC also improves over the length normalized variant for each subset.

5.2.3 Context-ignorant LM Probability

To understand how much of the unsupervised performance comes from context-independent LM biases, we analyze subsets where the correct answer is most/least likely without the context. We report the performance on the PIQA and ARC datasets in Table 5 and show that such biases indeed exist. The key takeaway is that the standard evaluation metrics may not give an accurate estimate of performance and that ALC provides more reliable estimates.

Finally, we report macro F1 scores for LM-Norm-Best and LM-Norm-Worst evaluation in Table 6 on PIQA and SWAG datasets. The results indicate that the datasets favours length normalization aware scoring irrespective of the context. When we measure the bias associated score, ALC generally performs better.

5.3 Expected Calibration Error

Given a score $S(O_k|C)$ for each choice O_k, we can compute a confidence estimate $\mathrm{conf}(O_k|C)$ as:

$$\mathrm{conf}(O_k|C) = \frac{e^{S(O_k|C)}}{\sum_{k' \in [K]} e^{S(O_{k'}|C)}} \tag{14}$$

Model	Accuracy (↑)	ECE (↓)
Length normalized	-0.46 (5.90)	-0.21 (0.13)
ALC (Unscaled)	+1.17 (6.23)	-0.07 (0.06)
ALC (BC)	+3.73 (3.86)	-0.09 (0.04)

Table 7: *Expected Calibration Error:* Mean (and standard deviation) of difference with the uncalibrated baseline in accuracy and ECE over different evaluation datasets are reported. ALC improves both ECE and accuracy. Please see Section 5.3 for more details.

Guo et al. (2017) compute expected calibration error (ECE) by partitioning N confidence predictions into R equal bins $B_r, r \in [1, R]$ and computing the weighted average of the absolute difference between the confidence and accuracy in each bin:

$$\mathrm{ECE} = \sum_{r=1}^{R} \frac{|B_r|}{N} |\mathrm{acc}(B_r) - \overline{\mathrm{conf}}(B_r)| \tag{15}$$

where $\mathrm{acc}()$ and $\overline{\mathrm{conf}}()$ measure the accuracy and mean confidence respectively in a bin. We set the number of bins to be 20.

We report the average difference in accuracy and ECE compared to the uncalibrated baseline across the evaluation datasets (except WinoGrande) in Table 7. When compared to the uncalibrated baseline, ALC provides gains in calibration error while also improving performance. Length-normalization also improves ECE, presumably by correcting for length bias. However, length-normalization does not improve performance on an average. The relative performance gains of ALC can be explained through the handling of additional biases beyond length bias.

6 Conclusion

We propose ALC (Answer-Level Calibration), an unsupervised method to improve performance of pretrained language models. We show that, when compared to existing baselines, ALC is more suitable for free-form multiple choice question answering, including commonsense reasoning tasks. We also show that popular datasets favor models which rely on easy cues for predictions, and that ALC provides more reliable estimates of model capabilities by getting rid of some of these biases.

References

Anil Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, entropy rates, and memory in language models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1089–1099. PMLR.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. 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)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4507–4518, Online. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 255–269, Online. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. 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)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Data	Size
COPA (Gordon et al., 2012)	100
COPA-test (Gordon et al., 2012)	499
CommonsenseQA (Talmor et al., 2019)	1221
MCTACO (Zhou et al., 2019)	454
SocialIQA (Sap et al., 2019)	1954
PIQA (Bisk et al., 2020)	1838
Winogrande (Sakaguchi et al., 2020)	1267
ALC (Easy) (Clark et al., 2018)	570
ALC (Challenge) (Clark et al., 2018)	299
SWAG (Zellers et al., 2018)	20000
DREAM (Sun et al., 2019)	2040
Hendrycks (Hendrycks et al., 2020)	14042

Table 8: Number of examples in the datasets used

Model	Size (Million parameters)
distilgpt2	82
gpt2-small	117
gpt2-medium	345
gpt2-large	774
gpt2-xl	1558

Table 9: Size of models used

A Appendix

A.1 Experimental Setup

We leverage the transformers library (Wolf et al., 2020) for accessing the LMs. All experiments were conducted using a single Nvidia GeForce GTX 1080 Ti Graphics Card. There was no training required. A typical experiment using gpt2-xl for CommonsenseQA task took around 15 minutes.

The model sizes are captured in Table 9. Size of evaluation datasets is captured in Table 8.

We used the nltk pos-tagger with the universal tagset for pos-tagging.

A.2 Datasets

COPA: The COPA dataset (Gordon et al., 2012) contains a premise associated with two alternatives where one has a more plausible causal connection with the premise. There are two types of examples, depending on whether the connection is of “effect" or “cause".

CommonsenseQA: The CommonsenseQA dataset (Talmor et al., 2019) contains common sense questions extracted from ConceptNet (Liu and Singh, 2004). The alternative choices are made challenging by selecting from related concepts in ConceptNet or through suggestions through crowdsourcing.

MCTACO: The MCTACO dataset (Zhou et al., 2019) contains common sense questions related to understanding of time. Difficult adversarial candidates are selected using BERT (Devlin et al., 2019) predictions.

SocialIQA: The SocialIQA dataset (Sap et al., 2019) contains questions about social interactions with crowdsourced answers.

PIQA: The PIQA dataset (Bisk et al., 2020) contains questions about common sense. The question corresponds to a goal derived from an instruction website and the answers were crowdsourced.

WinoGrande: The WinoGrande dataset (Sakaguchi et al., 2020) is based on the Winograd Schema Challenge (Levesque et al., 2012), where a pair of sentences differ in one or two words containing a referential ambiguity.

ARC: The ARC dataset (Clark et al., 2018) contains natural grade-school science questions. The authors provide Easy and Challenge splits. The Challenge version is created using examples where retrieval-based and word-occurrence based methods fail (Clark et al., 2018). The Easy version contains the remaining questions.

DREAM: DREAM (Dialogue-based REAding comprehension exaMination) (Sun et al., 2019) provides a benchmark for reading comprehension focusing on multi-turn multi-party dialog understanding.

SWAG: SWAG (Situations With Adversarial Generations) (Zellers et al., 2018) provides a large-scale dataset for grounded commonsense inference where different possible endings of a context are provided where the correct answer is derived from video captions while alternatives are adversarially generated.

Hendrycks: Hendrycks et al. (2020) provide a test suite containing 57 tasks to test the multitask accuracy of pretrained models. The tasks are broadly categorized into Humanities, STEM, Social Sciences and Other. We run our experiments on subsets associated with these categories.

A.3 Data Formatting

In this section, we provide the formatting used to convert task-specific examples into natural language prompts as used in our experiments. We first give examples of the **Context** (if any), the **Question** and **Choices** as present in the corresponding dataset, followed by the **Context input** and **Neutral input** as fed to the pretrained LM.

Model		COPA	COPA-test	CSQA	MCTACO	MCTACO-test	SocialIQA	PIQA	WG
Accuracy with roberta-large									
Uncalibrated		59.00	63.20	30.47	51.32	54.41	37.51	55.06	51.07
Length normalized		59.00	67.40	44.23	**54.85**	55.14	41.71	54.46	51.14
ALC	Unscaled	61.00	65.40	44.23	48.68	47.77	42.43	53.59	50.83
	TVD	**63.00**	65.60	44.47	53.74	55.60	40.63	56.64	51.30
	BC	**63.00**	**67.60**	**47.50**	54.63	**56.41**	**42.89**	**57.18**	**51.62**
1 shot									
Length normalized		-	3.19	12.97	-	1.53	4.41	-0.84	0.00
ALC	Unscaled	-	6.09	14.94	-	-7.33	5.65	1.61	0.14
	TVD	-	5.71	10.71	-	**2.65**	4.35	2.20	0.22
	BC	-	**6.29**	**17.60**	-	0.22	**5.82**	**2.78**	**0.25**
4 shot									
Length normalized		-	2.28	13.30	-	1.98	3.96	-1.14	-0.11
ALC	Unscaled	-	6.71	17.37	-	-7.26	5.69	3.63	0.34
	TVD	-	6.54	10.28	-	**2.81**	4.62	2.68	0.38
	BC	-	**7.14**	**17.74**	-	0.41	**6.27**	**3.86**	**0.39**

Table 10: *Standard evaluation results on unsupervised commonsense question answering tasks using RoBERTa-large*. As in Table 1, dev set accuracies (unless specified otherwise) are presented for ALC along with an unscaled version where the bias term is not scaled. The highest accuracies are marked in bold font. The trends are similar as observed in Table 1. Please see Section 5.1 for more details.

Model	Hendrycks			
	Humanities	STEM	Social sciences	Other
Baseline	27.28	25.11	32.03	32.62
Length normalized	29.26	**27.52**	32.00	34.16
ALC (Unscaled)	25.39	24.95	29.12	32.66
ALC (BC)	**31.40**	**27.52**	**33.81**	**35.68**

Table 11: LM-Best macro F1 evaluation on the Hendrycks Test using categories defined by Hendrycks et al. (2020).

- **CommonsenseQA**
  Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?
  Choices: (A) bank (B) library (C) department store (D) mall (E) new york
  Context input: Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? Answer:
  Neutral input: Answer:

- **MCTACO**
  Context: He layed down on the chair and pawed at her as she ran in a circle under it.
  Question: How long did he paw at her?
  Choices: (A) 2 minutes (B) 2 days (C) 3.5 hours (D) 1 day (E) 1.4 hours (F) 90 minutes (G) 7 hours (H) 7 days
  Context input: He layed down on the chair and pawed at her as she ran in a circle under it. Question: How long did he paw at her? Answer:
  Neutral input: Answer:

- **PIQA**
  Context: Remove soap scum from shower door.
  Choices: (A) Rub hard with bed sheets, then rinse. (B) Rub hard with dryer sheets, then rinse.
  Context input: Question: Remove soap scum from shower door. Answer:
  Neutral input: Answer:

- **ARC**
  Question: Which technology was developed most recently?
  Choices: (A) cellular telephone (B) television (C) refrigerator (D) airplane
  Context input: Question: Which technology was developed most recently? Answer:
  Neutral input: Answer:

- **COPA-effect**
  Context: The man turned on the faucet.
  Choices: (A) The toilet filled with water. (B)

Water flowed from the spout.
Context input: The man turned on the faucet, so
Neutral input: , so

- **COPA-cause**
  Context: The hamburger meat browned.
  Choices: (A) The cook froze it. (B) The cook grilled it.
  Context input: The hamburger meat browned, because
  Neutral input: , because

- **SocialIQA**
  The formatting follows Shwartz et al. (2020).
  Context: Tracy didn't go home that evening and resisted Riley's attacks.
  Question: What does Tracy need to do before this?
  Choices: (A) make a new plan (B) Go home and see Riley (C) Find somewhere to go
  Context input: Tracy didn't go home that evening and resisted Riley's attacks. Before, Tracy needed to
  Neutral input: Before, Tracy needed to

- **WinoGrande**
  Context: Sarah was a much better surgeon than Maria so _ always got the easier cases.
  Choices: (A) Sarah (B) Maria
  Context input: Sarah was a much better surgeon than Maria so
  Neutral input: so

- **DREAM**
  Context: W: I wish I knew the times of the trains to London. But our phone's out of order.
  M: Don't worry, Grandma. I'll find out for you on the Internet.
  W: Thank you!
  Question: What is the man going to do?
  Choices: (A) Go on the Internet. (B) Make a phone call. (C) Take a train trip.
  Context input: W: I wish I knew the times of the trains to London. But our phone's out of order.
  M: Don't worry, Grandma. I'll find out for you on the Internet.
  W: Thank you! Question: What is the man going to do? Answer:
  Neutral input: Question: What is the man going to do? Answer:

- **SWAG**
  Context: The person plays a song on the violin. The man
  Choices: (A) finishes the song and lowers the instrument. (B) hits the saxophone and demonstrates how to properly use the racquet. (C) ....
  Context input: The person plays a song on the violin. The man
  Neutral input: The man

- **Hendrycks**
  Question: If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
  Choices: (A) 28 (B) 21 (C) 40 (D) 30
  Context input: Question: If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps? Answer:
  Neutral input: Answer:

A.4 Note on Comparison with Self-talk

While it seems surprising that the self-talk results in Table 1 are generally lower than the uncalibrated baseline, we note that we haven't underestimated the performance of self-talk. Self-talk performance was obtained using the official repository of the project and the results align well with those reported in the original work. What has changed is the performance of the baseline, which is higher here (which in turn shows the significane of the numbers reported in this work). We note two differences with respect to the self-talk repository. First, self-talk uses a length-normalized baseline, while we evaluate both uncalibrated and length-normalized baselines. Second, there is a bug in the self-talk repository regarding calculating baseline performance, also noted in a GitHub issue[7].

A.5 Additional Results

We show results across commonsense reasoning datasets for bias-associated F1 scores in Table 12.

[7] https://github.com/vered1986/self_talk/issues/1

Model	Acc	$\mathrm{F1}_{\mathrm{Shortest}}$	$\mathrm{F1}_{\mathrm{Longest}}$	$\mathrm{F1}_{\mathrm{POS}}$	$\mathrm{F1}_{\mathrm{LM-Best}}$	$\mathrm{F1}_{\mathrm{LM-Worst}}$
COPA-dev						
Uncalibrated	72	71.9	71.96	71.36	68.81	68.81
Length normalized	68	67.68	67.68	68.1	66.04	66.04
ALC (Unscaled)	70	69.81	69.89	68.2	69.95	69.95
ALC (BC)	**73**	**72.78**	**72.78**	**71.64**	**72.67**	**72.67**
COPA-dev						
Uncalibrated	74.2	73.95	74.04	74.15	71.3	71.3
Length normalized	72.8	72.73	72.66	72.45	71.31	71.31
ALC (Unscaled)	79.2	79.18	79.2	78.2	79.18	79.18
ALC (BC)	**80**	**79.97**	**80**	**78.96**	**79.78**	**79.78**
CSQA						
Uncalibrated	37.18	36.33	35.67	35.48	36.9	33.1
Length normalized	33.82	30.01	34.94	34.92	29.33	34.68
ALC (Unscaled)	47.91	43.9	48.99	49.84	43.1	48.9
ALC (BC)	**49.71**	**45.95**	**49.96**	**51.67**	**45.97**	**50.47**
MCTACO						
Uncalibrated	61.89	61.24	64.65	63.16	59.03	66.77
Length normalized	55.73	58.01	57.76	55.45	57.28	61.87
ALC (Unscaled)	57.05	59.51	60.07	57.4	59.39	61.49
ALC (BC)	**64.76**	**65.89**	**67.52**	**66.07**	**64.99**	**69.64**
SocialIQA						
Uncalibrated	40.53	40.93	34.4	35.77	37.77	29.25
Length normalized	41.35	32.58	40.83	41.63	39.19	42.09
ALC (Unscaled)	42.68	40.41	43.41	42.56	40.14	43.42
ALC (BC)	**45.14**	**44.37**	**45.68**	**45.04**	**44.69**	**45.82**
PIQA						
Uncalibrated	70.67	69.9	69.79	74.3	59.83	59.83
Length normalized	71.33	**71.33**	**71.33**	**74.79**	65.58	65.58
ALC (Unscaled)	59.96	59.95	59.95	56	59.51	59.51
ALC (BC)	**70.78**	70.39	70.33	73.95	**66.82**	**66.82**

Table 12: *Overall bias associated evaluation results:* We present bias-associated F1 scores for each attribute considered. We note that ALC consistently performs better or as competitive with the baselines. Please see Section 5.2 for details.