RBCorr: Response Bias Correction in Language Models

Om B. Bhatt, Anna A. Ivanova
om.bhatt@uci.edu, a.ivanova@gatech.edu
Language, Intelligence, & Thought (LIT) Lab, Georgia Tech

ACL 2026 SAN DIEGO JULY 2-7

Background

- Language models (LMs) often show *option preference* in closed-choice questions, a form of "response bias".
- Response bias can be detrimental towards assessing LMs' true reasoning abilities and potentially dangerous for models deployed IRL.
- Recent works have attempted to use token log-probabilities (**LogProbs**) as a cost-effective way to identify and correct response bias.
- However, bias is not comprehensively tracked, corrections focus only on accuracy improvement, and relationships between bias and other aspects of the evaluation setup are not studied.

We propose a simple correction method, RBCorr, and test it on 12 open-weight LMs across three question-types. RBCorr effectively mitigates bias while preserving or improving model performance.

RBCorr yields comparable or higher gains than two existing LogProbs-based methods and shows variable efficacy dependent on other aspects of the evaluation setup.

Setup

We derive **base** LM responses via single-token LogProbs:

Q: Sally pays salary to Harry. Is Sally Harry's boss?

$P(Q|Yes/\_Yes) = -4.675$
$P(Q|No/\_No) = -5.179$

Model Answer: **YES**

❖ We test LMs using three question-types across 10 datasets:
1. **Yes-No (Yes / No):** ARITH, BABI, COMPS, EWOK
2. **Entailment (0 / 1 / 2):** MNLI, SNLI
3. **4-Choice (A / B / C / D):** MMLU – STEM, Humanities, Social Sci., Others

Each dataset has 1200 *class-balanced* items to reveal option preference.

❖ We test 12 models split across three families (instruct/non-instruct for each):
1. Falcon-3-(3B/10B) **2.** Gemma-3-(12B/27B) **3.** Llama-3.1-(8B/70B)

❖ We explore response behavior and bias correction efficacy in relation to:
1. **Prompt Complexity:** Using *zero-shot*, *instruction-only*, and *few-shot* prompts.
2. **Question Type:** Test different option-space sizes and knowledge domains.
3. **Model Family:** Test LMs with shared architecture and training procedures.

Bias Metrics

Total Variation Distance (TVD) and **Relative Standard Deviation** (RSD):

$$\textbf{TVD}(G, M) = \frac{1}{2} \sum_{x \in X} |G(x) - M(x)|$$

(G: ground truth dist., M: model response dist.)

$$\textbf{RSD} = \frac{\sqrt{\frac{1}{|X|} \sum_{i=1}^{|X|} (acc_i - acc)^2}}{acc}$$

(X: option space, acc: mean acc over all labels)

- **TVD** more sensitive to *overall* dist. shifts, **RSD** more sensitive to *per-class* dist. shifts
- Both metrics are relative, i.e., lower value after correction implies bias reduction.

RBCorr Method

Steps:

1. Extract LogProbs values for each option label for all items in the dataset.
2. Sample a small class-balanced set of questions from the dataset (i.e., *calibration set*).
3. Calculate the per-label mean of LogProbs values using the calibration set (i.e., *correction term*).
4. Subtract the label-specific correction term from the LogProbs values for all questions *outside* of the calibration set, i.e., the *evaluation set*.

➢ To achieve complete evaluation coverage, RBCorr is applied in a ***k-fold*** manner using separately-sampled calibration sets.

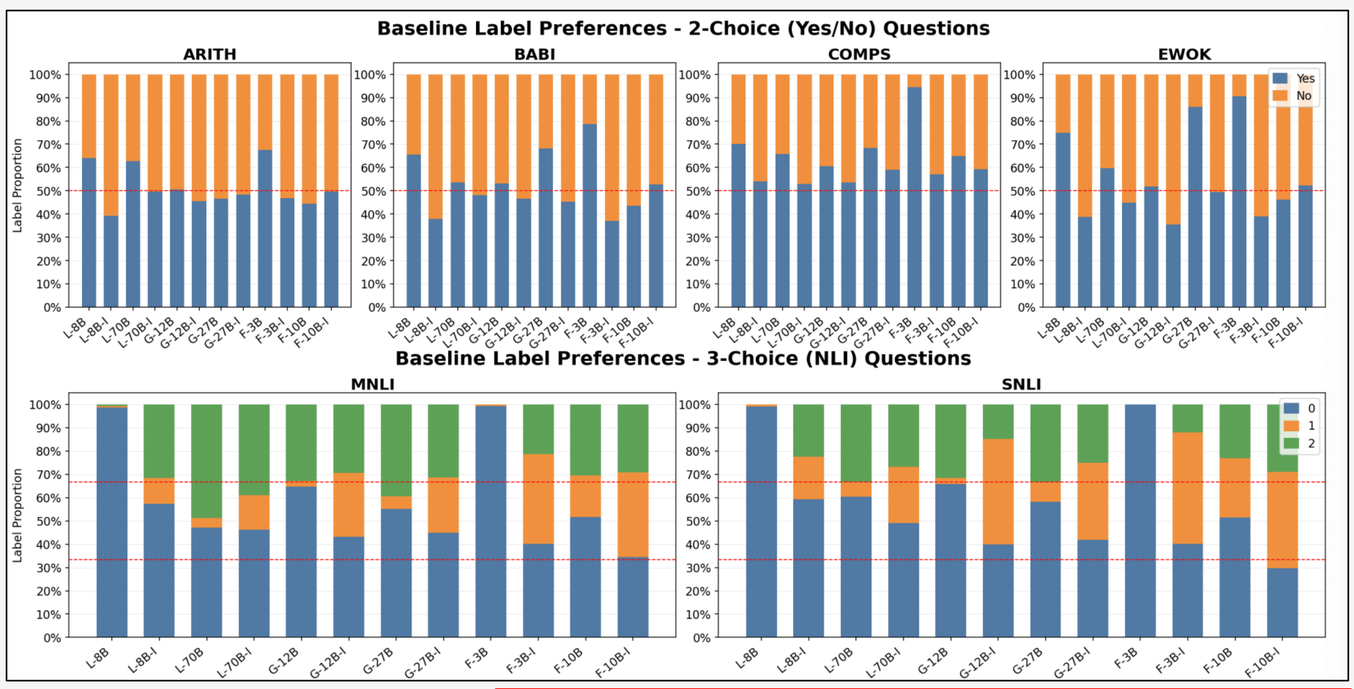
➢ We test calibration sets of size {24, 60, 120, 180, 240} at k = 5; we show RBCorr is **unaffected** by set size.

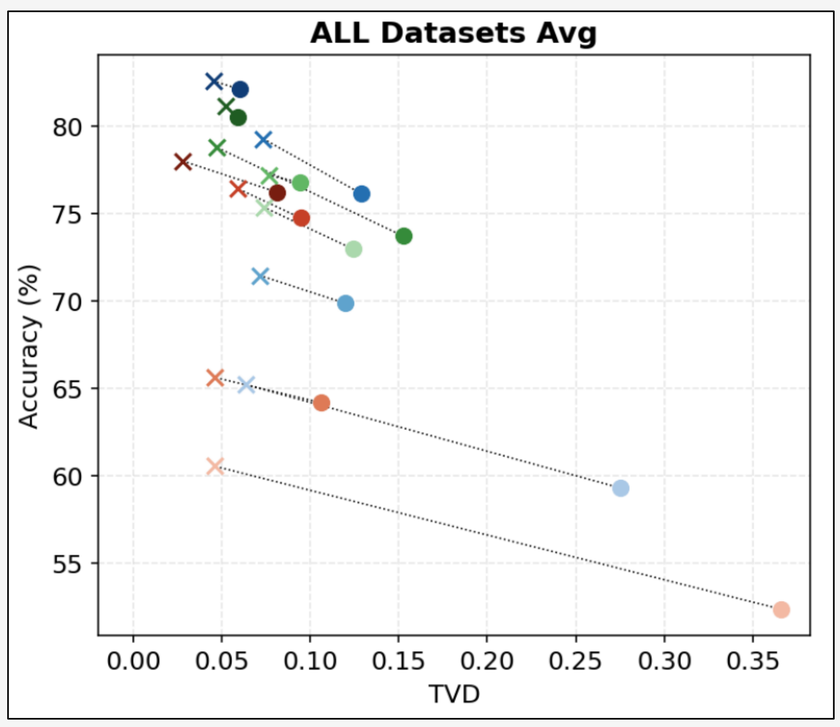
Existing Methods

We compare RBCorr to two existing methods:

1. **Contextual Calibration (CC):** Collecting "content-free" label LogProbs values by substituting strings like "N/A"/"[MASK]" into question templates and using their means as correction terms.
2. **Batch Calibration (BC):** Similar to RBCorr, uses in-dataset samples to estimate a correction term. However, BC does not enforce class balance in picking calibration sets (thus not requiring ground truth labels), and applies correction terms to the calibration set itself, thus dissolving train-test separability.

Results





▲ High-accuracy, low-bias LMs should place at the **top-left** of scatterplot ▲

- **Instruction-tuned** and **bigger-sized models** generally show **less-biased behavior** compared to smaller and non-instruct counterparts.
- **Few-shot** prompting generally reduces bias compared to zero-shot or instruction-only prompts.

Method	Acc. (%) (↑)	TVD (↓)	RSD (↓)
RBCorr vs CC	+3.00***	−0.0956***	−0.1404**
RBCorr vs BC	**+0.15****	+0.0003 **n.s.**	−0.0008 **n.s.**
BC vs CC	+2.85***	−0.0959***	−0.1396***

- **RBCorr yields the highest accuracy gains** compared to BC and CC!
- Bias reduction efficacy difference between RBCorr and BC is non-significant.
- RBCorr performance is similar to BC overall but offers slightly higher accuracy gain and train-test separability to improve evaluation integrity.

Future Work

1. It could be possible to do an initial (post-hoc) correction and **use the corrected set of LogProbs as a tuning set** to inherently de-bias the model's outputs.
2. We could try using the **outputs of a strong 3rd-party model as a ground-truth proxy**, thus removing the requirement for ground-truth labels.
3. We could try **using more sophisticated correction-term calculations**, e.g., stratifying items via LogProbs entropy (confidence) and calculating separate sets of correction terms.
4. Mechanistic **interpretability methods** could be used to identify bias at the layer-level rather than the output-token level and finding ways to correct bias there.

References

Bias Metrics:
- Reif, Y., & Schwartz, R. (2024). Beyond Performance: Quantifying and Mitigating Label Bias in LLMs (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.02743Reif, Y., & Schwartz, R. (2024). Beyond Performance: Quantifying and Mitigating Label Bias in LLMs (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.02743

Contextual Calibration:
- Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2102.09690

Batch Calibration:
- Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., & Roy, S. (2023). Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2309.17249