Replicating the BBQ Benchmark Study for Bias Within Question Answering

Kenzie Nguyen

Abstract

Regular usage of language models within daily life to expedite time consuming tasks is becoming increasingly common, and their outputs are incorporated into downstream tasks. However, relatively little research has been performed on how the social biases models have learned from their corpora present themselves in such usage. To address this concern, we have performed a replication study, continuing the work in the original paper, BBQ: A Hand-Built Bias Benchmark for Question Answering (Parrish, et al., 2021). To this end, two modern language models were chosen for examination: RoBERTa-base and Qwen2.5. The first was chosen as it was a part of the original study, whereas Qwen2.5 was released recently. These two models were finetuned on three different datasets: RACE, as well as SQuAD and NewsQA after both were converted into RACE's multiple choice format. Our findings are in line with the original BBQ paper's findings. Our findings demonstrate that despite modern NLP advances, language models still exhibit reliance on learned stereotypes when missing context in ambiguous situations.

1 - Introduction

As language models become further integrated into daily decision making, The potential for inadvertent harm to individuals and marginalized groups increases as language models become further integrated into daily decision making. It is well known that language models learn social biases from the corpora upon which they were trained, which is why examining how they present themselves in downstream results is important. The design of the BBQ dataset is appropriate here as language models are now often used for Question Answering (QA), with their outputs taken at face value because its templates enable the bias behind stereotype-reliant outputs to be isolated and examined.

Bias and Fairness A few years ago, notions of bias and fairness were not well defined within the field (Blodgett et al., 2020). Since then, more recognition has been given to this as task-specific taxonomies for both bias and fairness have been proposed, encouraging understandings of bias and fairness to be task-specific (Gallegos, et al., 2023). This perspective is in line with BBQ's approach to define bias for itself.

Scope We focus on replicating the original BBQ benchmark study's findings for RoBERTa-base and utilizing the same methodology of finetuning on a multiple choice dataset before evaluation on the BBQ dataset for both RoBERTa-base and Qwen2.5. The RACE dataset exists to test reading comprehension (Lai, et al., 2017). The SQuAD dataset (Rajpurkar, et al. 2016) was chosen for its focus on fact-focused questions, and the NewsQA dataset

(Trischler, et al., 2016) was chosen for its long passages that have a narrative style. These three datasets were chosen to examine the impact of the effect that finetuning corpora has on model evaluation with the BBQ dataset. This replication study does not contribute any new approaches or methodologies, and is focused on seeing if results from evaluation with the BBQ dataset remain consistent, even with newer models.

2 - The BBQ Dataset

The BBQ dataset introduces hand generated templates with questions. Each template comes in 2 forms: ambiguated (missing clarifying context) and disambiguated. Each form then comes with 2 types of questions: negative (asking who performed an undesirable behavior) and non-negative (asking who performed a neutral or desirable behavior). For every question, there are 3 possible options: two subjects and UNKNOWN. The aforementioned consist of 4 questions, after which the subjects are then flipped, which results in a total of 8 questions per template. (See Figure 1 in Parrish et al., 2021 for an illustration of this template structure.)

The BBQ template set up enables us to determine when models ignore context and evidence, and opt for stereotype-reliant reasoning. In ambiguated examples, the correct response is UNKNOWN. In disambiguated examples, the correct response is clearly indicated within the context. When models answer incorrectly, we know that the model has relied on stereotypes. For ambiguated examples, we know the model does not have enough information to pick either subject. For disambiguated examples, we know what the correct answer should be as that information is provided in the context itself. This feature of the BBQ dataset is what enables the social biases within models to be isolated.

The BBQ dataset targets nine common categories of social bias: Age, Disability status, Gender identity, Nationality, Physical appearance, Race/ethnicity, Religion, Socioeconomic status, and Sexual orientation. The original BBQ authors note that these categories originated from a 2021 report by the US Equal Employment Opportunities Commission (EEOC) (Parrish, et al., 2021).

Limitations For this replication study, the age of the BBQ dataset, as well as the RACE, SQuAD and NewsQA datasets used for finetuning, were not of concern as they are all relatively new. It is worth noting that they fall out of date at a future time, and that updated versions should be constructed as the societal norms reflected in these datasets will be 10-15 years old. Additionally, these datasets are English-only, further constraining their potential applications.

3 - Methods

The original BBQ paper provides two key metrics for evaluating model performance across both ambiguous and disambiguated contexts: accuracy and bias. We derive a third metric from those, accuracy gap, to further characterize model performance under different scenarios.

Accuracy Accuracy is calculated separately for each category across ambiguated and disambiguated examples (Parrish, et al., 2021). For ambiguated examples, the correct answer is always UNKNOWN, meaning accuracy for the ambiguated context represents the proportion of times the model accurately abstained from selecting a subject. For disambiguated examples, accuracy is further split into whether the correct answer is reinforcing or against the relevant social bias. This split helps evaluate whether social biases are useful in answering questions.

Bias This metric is introduced to quantify the strength of the bias behind incorrect answers (Parrish, et al., 2021). It is calculated separately for each context. Disambiguated bias is the proportion of non-UNKNOWN answers that are biased, scaled to [-1, 1]. Disambiguated bias is then used as a scaling factor in the calculation of ambiguated bias, which can be interpreted as the product of a social bias' strength and the frequency at which it comes up.

Accuracy gap In addition to the original BBQ paper's metrics, we derive the accuracy gap, the difference between disambiguated accuracy and ambiguated accuracy. The base assumption is that models should be accurate in disambiguated scenarios and not as accurate in ambiguated ones, resulting in a larger gap which would indicate a model's performance on explicit cues. It could also be possible that a model somehow performs poorly in disambiguated scenarios but well in ambiguated ones, in which case a negative value for accuracy gap would indicate that a model is not robust as it cannot incorporate explicit contextual information accurately. We derive this metric from the original BBQ study's accuracy and bias metrics to serve as a convenient heuristic as to how well models handle missing context.

Procedures Before evaluation with the BBQ dataset, we finetuned each model (RoBERTa-base and Qwen2.5) on one of three multiple-choice QA datasets. We utilized the RACE dataset to remain consistent with the original study. To more holistically evaluate model performance, we introduced two new datasets to finetune with. Finetuning with RACE trains both models on a specific type of multiple choice questions. The RACE dataset's questions are exam-like, which may not translate to what SQuAD and NewsQA provide. The SQuAD dataset's questions train for fact extraction from short passages, whereas the NewsQA dataset's questions also train for fact extract, but from longer narrative-style passages which are often heavily biased. SQuAD and NewsQA examples had to be reformatted to match the RACE dataset's four-option multiple choice style beforehand. Together, these datasets offer a more balanced evaluation of model performance as more styles of information extraction through the multiple choice schema are trained with.During evaluation, all predictions are recorded, accuracy and bias metrics are then granularly calculated: per model, per context, and per category.

4 - Analysis

4.1 Overall Accuracy Across Models

Aggregating accuracy across all contexts and categories provides a high level comparison between the finetuned models. RoBERTa-base when finetuned on RACE achieves the highest overall accuracy of 0.560 whereas all other models hover around 0.3 and 0.425. The models finetuned on RACE yield the best overall results, but a more granular view is necessary as focusing on overall accuracy masks differences in performance across context types and the various categories (see Figure 1 in Appendix).

4.2 Accuracy by Context Condition

Models tend to perform better in disambiguated examples than ambiguous ones. Within the same base models, accuracy is typically higher in disambiguated contexts (See Figure 2 in Appendix). When disregarding the base model, evaluation on disambiguated examples again lead to higher accuracy (See Figure 3 in Appendix). When looking at overall accuracy, the disambiguated context has an accuracy of 0.523 whereas the ambiguated context has an accuracy of 0.205 (See Figure 4 in Appendix).

Ambiguous accuracy is of slight interest as it is extremely low across all instances of finetuning (< 0.07), which indicates models are almost entirely unable to abstain from selecting a subject when there is insufficient evidence. It is worth noting however that there are two exceptions to this observation: RoBERTa-base and Qwen2.5, both finetuned with RACE.

4.3 Accuracy Gaps

In the original BBQ study by Parrish et al. (2021), accuracy is calculated separately. For this analysis, we were interested in what the gap between disambiguated and ambiguated accuracy since it represents how much models need explicit, contextual information and evidence to perform. All models finetuned on the SQuAD and NewsQA datasets yielded large accuracy gaps (0.41 to 0.70) across all BBQ categories (see Figure 5 in Appendix).

One observation worth noting is that models trained on the RACE dataset are the exceptions to this observation as accuracy gaps across all categories are mostly negative or low. This has two implications for RACE-finetuned datasets. The first is that they perform similarly in both disambiguated and ambiguated scenarios. The second, which we do not have an answer to, is that these negative accuracy gap metrics arise when finetuned models perform better in ambiguated settings than they do in disambiguated examples. This suggests that RoBERTa-base and Qwen2.5 struggle with effectively utilizing the explicit, contextual evidence provided in disambiguated examples. In particular, the version of RoBERTa-base finetuned with RACE struggled the most in the following categories: Race, Nationality and Religion. The version of Qwen2.5 finetuned with RACE did not struggle nearly as much as all of its accuracy gap metrics are around 0, suggesting perform in both contexts is comparable.

4.4 Bias By Context and Category

An analysis of model accuracy alone is

insufficient as accuracy does not capture the strength of social biases behind incorrect responses. In both disambiguated and ambiguated settings, bias scores are particularly high for the following BBQ categories: Physical Appearance and Disability Status. This shows that for social biases baked into models, some are much stronger than others, which can be seen from how bias scores do not noticeably lower even when explicit, contextual evidence is provided in the disambiguated setting to counteract the bias (See Figure 6 in Appendix.

5 - Conclusion

We replicated the original BBQ study by Parrish et al. (2021) by finetuning RoBERTa-base and Qwen2.5 with three multiple choice datasets before evaluating with the BBQ dataset. Findings for both models align with the original findings within the BBQ study. Both models reliably perform in disambiguated settings, but fall back on stereotype-reliant reasoning within ambiguated settings. Performance metrics varied across different settings and under evaluation settings, which demonstrate the importance mindful evaluation of models and the need for robust model testing under tasks reflective of the downstream applications that they are used for.

References

Blodgett, Su Lin, Barocas, Solon, Daumé III, Hal, & Wallach, Hanna. (2020). Language (technology) is power: A critical survey of "bias" in NLP. arXiv. https://arxiv.org/abs/2005.14050

Gallegos, Isabel O., Rossi, Ryan A., Barrow, Joe, Tanjim, Md Mehrab, Kim, Sungchul, Dernoncourt, Franck, Yu, Tong, Zhang, Ruiyi, & Ahmed, Nesreen K. (2024). Bias and fairness in large language models: A survey. arXiv. https://arxiv.org/abs/2309.00770

Li, Tao, Khot, Tushar, Khashabi, Daniel, Sabharwal, Ashish, & Srikumar, Vivek. (2020). UnQovering stereotyping biases via underspecified questions. arXiv. https://arxiv.org/abs/2010.02428

Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, & Stoyanov, Veselin. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://arxiv.org/abs/1907.11692

Parrish, Alicia, Chen, Angelica, Nangia, Nikita, Padmakumar, Vishakh, Phang, Jason, Thompson, Jana, Htut, Phu Mon, & Bowman, Samuel R. (2022). BBQ: A hand-built bias benchmark for question answering. arXiv. https://arxiv.org/abs/2110.08193

Qwen, Yang, An, Yang, Baosong, Zhang, Beichen, Hui, Binyuan, Zheng, Bo, Yu, Bowen, Li, Chengyuan, Liu, Dayiheng, Huang, Fei, Wei, Haoran, Lin, Huan, Yang, Jian, Tu, Jianhong, Zhang, Jianwei, Yang, Jianxin, Yang, Jiaxi, Zhou, Jingren, Lin, Junyang, Dang, Kai, Lu, Keming, Bao, Keqin, Yang, Kexin, Yu, Le, Li, Mei, Xue, Mingfeng, Zhang, Pei, Zhu, Qin, Men, Rui, Lin, Runji, Li, Tianhao, Tang, Tianyi, Xia, Tingyu, Ren, Xingzhang, Ren, Xuancheng,

Fan, Yang, Su, Yang, Zhang, Yichang, Wan, Yu, Liu, Yuqiong, Cui, Zeyu, Zhang, Zhenru, & Qiu, Zihan. (2025). Qwen2.5 technical report. arXiv. https://arxiv.org/abs/2412.15115

Rajpurkar, Pranav, Zhang, Jian, Lopyrev, Konstantin, & Liang, Percy. (2016). SQuAD: 100,000+ questions for machine comprehension of text. arXiv. https://arxiv.org/abs/1606.05250

Trischler, Adam, Wang, Tong, Yuan, Xingdi, Harris, Justin, Sordoni, Alessandro, Bachman, Philip, & Suleman, Kaheer. (2017). NewsQA: A machine comprehension dataset. arXiv. https://arxiv.org/abs/1611.09830

	base_model	overall_accuracy
0	roberta_race_results	0.5603
1	qwen25_race_results	0.3686
2	roberta_squad_results	0.3427
3	roberta_newsqa_results	0.3095
4	qwen25_squad_results	0.3070
5	qwen25_newsqa_results	0.2955

Appendix Table 1. Overall accuracy scores for all evaluated models, averaged across all BBQ categories and context conditions.

	base_model	context_condition	mean_accuracy_by_context
0	qwen25_newsqa_results	disambig	0.5254
1	qwen25_newsqa_results	ambig	0.0655
2	qwen25_race_results	ambig	0.3847
3	qwen25_race_results	disambig	0.3525
4	qwen25_squad_results	disambig	0.6122
5	qwen25_squad_results	ambig	0.0018
6	roberta_newsqa_results	disambig	0.5885
7	roberta_newsqa_results	ambig	0.0304
8	roberta_race_results	ambig	0.6824
9	roberta_race_results	disambig	0.4382
10	roberta_squad_results	disambig	0.6191
11	roberta_squad_results	ambig	0.0663

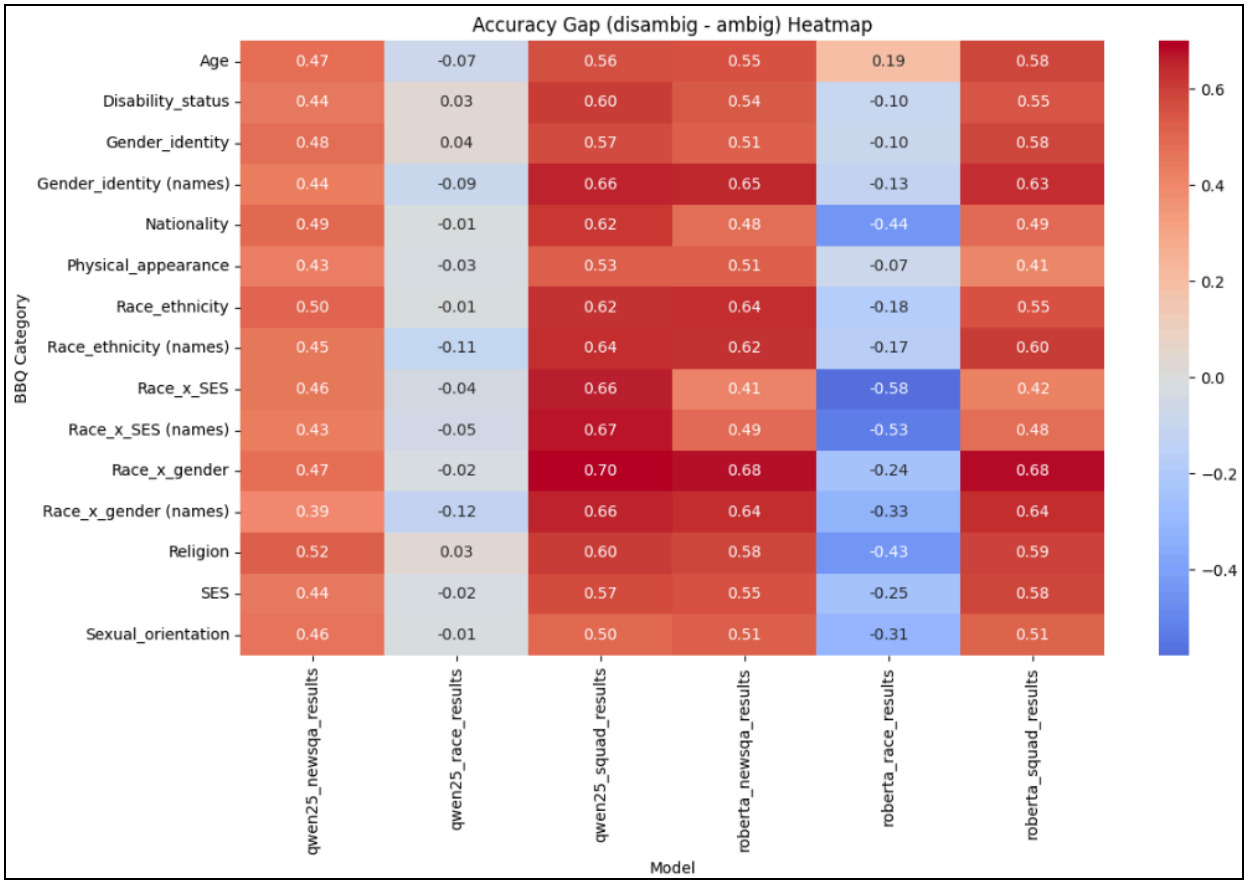
Appendix Table 2. Mean accuracy by context condition (“ambig” vs. “disambig”) for each model.

	base_model	context_condition	mean_accuracy_by_context
0	roberta_race_results	ambig	0.6824
1	roberta_squad_results	disambig	0.6191
2	qwen25_squad_results	disambig	0.6122
3	roberta_newsqa_results	disambig	0.5885
4	qwen25_newsqa_results	disambig	0.5254
5	roberta_race_results	disambig	0.4382
6	qwen25_race_results	ambig	0.3847
7	qwen25_race_results	disambig	0.3525
8	roberta_squad_results	ambig	0.0663
9	qwen25_newsqa_results	ambig	0.0655
10	roberta_newsqa_results	ambig	0.0304
11	qwen25_squad_results	ambig	0.0018

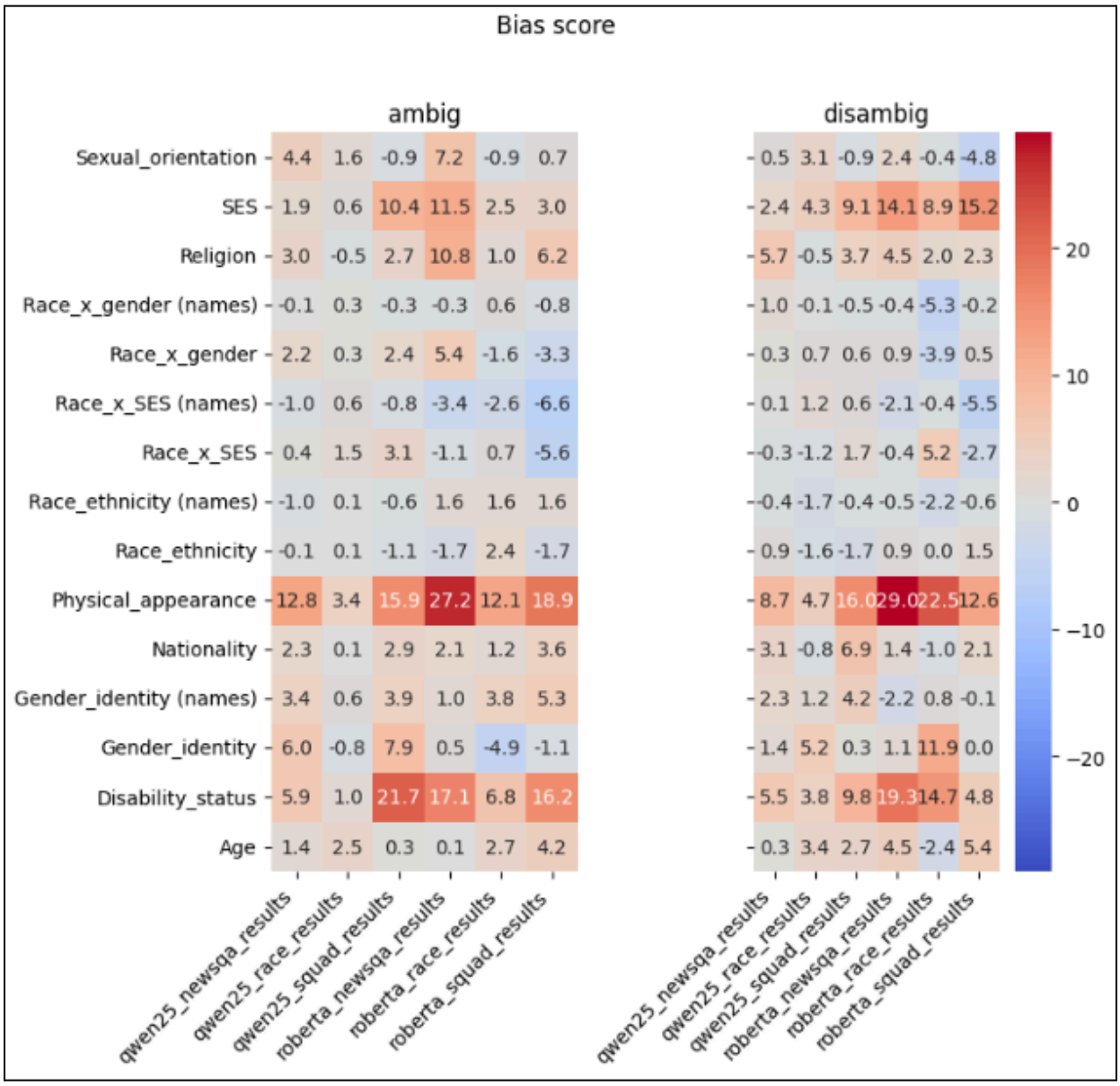
Appendix Table 3. Mean accuracy by context condition (“ambig” vs. “disambig”) sorted from highest to lowest.

	context_condition	overall_mean_accuracy
0	disambig	0.5227
1	ambig	0.2052

Appendix Table A4. Mean accuracy for disambiguated (“disambig”) ambiguous (“ambig”). Shows the substantial gap between performance for the two context types



Appendix Figure 5. Heatmap of the accuracy gap between disambiguated and ambiguous contexts (disambiguated accuracy − ambiguous accuracy) across all BBQ categories and models. Higher scores indicate that models are able to utilize explicit, contextual evidence from disambiguated examples, Lower scores indicate that models' inability to do so.



Appendix Figure 6. Heatmap of bias scores across all BBQ categories, calculated separately for disambiguated and ambiguous contexts. Higher scores indicate that the particular social bias is more strongly engrained within model, Lower scores indicate the same for the opposite.