Estimating and Correcting Yes-No Bias in Language Models

Om Bhatt (om.bhatt@gatech.edu)
College of Computing
Georgia Institute of Technology

Anna A. Ivanova (a.ivanova@gatech.edu)
School of Psychology
Georgia Institute of Technology

Abstract

When presented with a yes-no question, humans tend to say 'yes' regardless of the ground truth. This 'yes-bias' can be attributed either to the social pressure to agree with an interlocutor or simply to the tendency to mimic the distribution of the input data. Here, we estimate 'yes-no' response bias in language models (LMs), with the goal of distinguishing the two theories, and explore two strategies for bias correction. We develop two yes-no question datasets derived from existing world knowledge datasets, and test 16 open-weight LMs. We find that LMs often show response bias on yes-no questions, but that it is highly variable, deviating from bias observed in humans. We further present a novel bias correction method, which eliminates bias and improves model performance. Evidence of non-humanlike response bias in LMs informs us on the source of yes-bias in humans, and the efficacy of our bias correction method holds promise for LM evaluation. [1]

Keywords: language models; question-answering; bias correction

Introduction

Bias has been a long-standing area of investigation in cognitive science, and has experienced a recent influx of analytical paradigms with the rise of machine learning research. Response bias, in particular, is a topic of interest shared by both cognitive scientists (Fazio & Olson, 2003; McGrath, Mitchell, Kim, & Hough, 2010; Paulhus, 1991) and ML scientists (Dai et al., 2024; Zou, Mubin, Alnajjar, & Ali, 2024, among others). If an agent (human or model) answers a question or responds to a survey under the influence of systematic bias, it can lead to a false assessment of the agent's underlying knowledge and capabilities. Thus, for both humans and models, ensuring that a response to a question is a fair representation of the underlying judgment requires the identification, measurement, and correction of response biases.

One of the simplest types of questions presented in question-answering tasks are binary choice, yes-no questions. Decades of human behavioral research have shown evidence of an acquiescence bias (henceforth 'yes-bias'). In adults, yes-bias is most commonly studied in self-reported personality tests, where it holds major implications for personality research (Danner, Aichholzer, & Rammstedt, 2015), marketing research (Steinmetz & Posten, 2020), and possibly even measuring political efficacy (Wright, 1975). In children, yes-bias has not only been established (Peterson, Dowden, & Tobin, 1999), but further research has revealed its presence with varying ages, cultures, and languages (Heather Fritzley & Lee, 2003; Mehrani & Peterson, 2017; Okanda & Itakura, 2008).

The most common explanation for yes-bias is a social one: saying 'yes' has been historically attributed to the garnering of social desirability (Furnham, 1986; van de Mortel, 2008; Randall & Fernandes, 1991), and secondarily as a pressure of authoritarianism in various scenarios (Bass, 1955; Fox, Payne, Priest, & Philliber, 1977). It may also simply be a product of avoiding social confrontation. However, another possible explanation is a distributional one: children, and possibly even adults, might be saying 'yes' more often because it is a more common distributional pattern in their language input, causing it to become a default response.

We can test the distributional hypothesis of yes-bias origin by examining bias for either 'yes' or 'no' (henceforth 'yes-no bias') in language models (LMs). Existing studies show some evidence of response biases in large LMs (LLMs), but work surrounding acquiescence is both scarce and conflicting: (Tjuatja, Chen, Wu, Talwalkwar, & Neubig, 2024) show slight acquiescence in LLMs with heavily suggestive question re-wording, indicating a lack of inherent yes-bias, while (Schoenegger, Tuminauskaite, Park, Bastos, & Tetlock, 2024) establish yes-bias using LLM ensemble predictions. (Salecha et al., 2024) show evidence of human-like social desirability in LLMs without traceability to acquiescence. Thus, the evidence and nature of yes-bias in LMs is inconclusive, and more work is needed on both the investigative front to provide bias measurement/correction methods, as well as the evaluative front to provide high-quality benchmarking datasets.

It is also important to characterize LM bias on yes-no questions to be able to accurately evaluate their capabilities. Similar to how task demands might overwhelm smaller models and hinder successful task performance (Hu & Frank, 2024), response biases might mask underlying LLM knowledge. If such bias exists and is successfully corrected, LMs may show better performance on diverse tasks and thus serve as better cognitive models of language-based learning. Currently, the question of whether a bias towards yes-no questions exists in LMs remains open, as well as the prospect to correct for it.

This paper aims to answer the following questions: (1) whether a systematic yes-no bias exists in LMs, (2) if so, whether it can be corrected at the token distributional level (specifically via LogProbs manipulation), and (3) how the

[1] Code: https://github.com/ombbhatt/logprobs-bias-correct

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

corrections affect model accuracy. To do so, we convert two pre-existing world knowledge datasets into yes-no question format and measure the yes-no bias in LMs across four model families. We then measure the efficacy of two bias correction methods, one being model-inherent ('Generic') and the other being dataset-specific ('Specific'). Our results suggest that models do exhibit response bias, although their behavior does not generalize to a consistent yes-bias. We show that the model biases are dependent on a number of factors, but can be effectively corrected using the dataset-specific approach, which holds theoretical parallels to human-based bias mitigation methods.

Methods

Datasets

We adapt two world-knowledge datasets, COMPS (Misra, Rayz, & Ettinger, 2023) and EWoK (Ivanova et al., 2024), and convert them into two yes-no question datasets, COMPS-YNQ and EWoK-YNQ (YNQ → 'Yes-No Questions'). Our goal was to leverage datasets that are cognitively motivated, simple, and class-balanced (equal number of Yes and No gold label responses).

COMPS-YNQ COMPS consists of 'properties' (e.g., `basks in the sun`) and 'concepts' (e.g., `iguana`, `trolley`, etc.) that construct minimal pair sentences (`{An iguana / A trolley} basks in the sun.`). The minimal pairing allows for systematic conversion into yes-no questions (`Does {an iguana / a trolley} bask in the sun?`). This data set is intended primarily to test the model's ability to attribute properties to concepts based on observed experience, i.e., the retrieval of its pre-trained knowledge. We specifically chose the subset of negative concepts generated via 'random' sampling from the original dataset to construct COMPS-YNQ, since models reported the highest distinguishing capability between positive and 'random' negative concepts on average (Misra et al., 2023). The dataset contains a total of 7,184 questions, of which we randomly sampled **2,100** questions (ensuring class balance).

EWoK-YNQ EWoK introduces contextual plausibility as an additional factor affecting the concept-property matching process (Sinha et al., 2022). The original dataset contrasts opposing concepts (e.g., `help, hinder`) as minimal pairs of 'target' sentences (e.g., `Chao is {helping / hindering} Yan`), and the models are tested on their ability to match target sentences with corresponding pairs of 'context' sentences (e.g., `Chao is making Yan's job {easier / harder}.`). We convert the target sentences into yes-no questions and append them to the context sentences (`Chao is making Yan's job {easier / harder}. Is Chao {helping / hindering} Yan?`). EWoK-YNQ consists of **2,056** context-question input sequences across 11 domains (social interactions, physical relations, agent properties, etc.). Each question is fully answerable based on the single-sentence context, and the model must not only remember the immediate context association, but also judge its plausibility based on pre-training knowledge to answer the question correctly.

Models

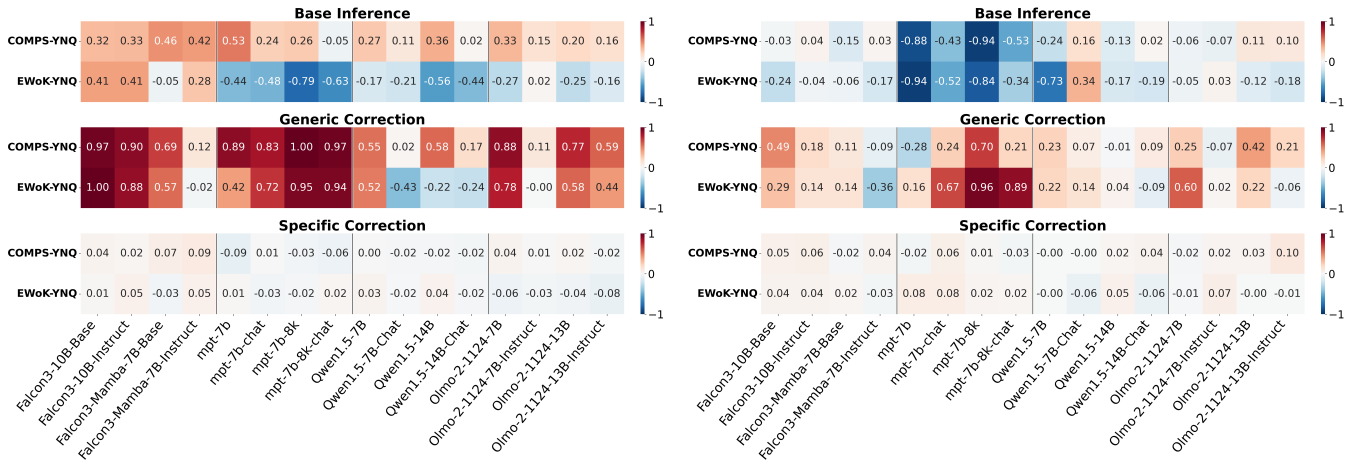
We selected 16 total LMs for evaluation on COMPS-YNQ and EWoK-YNQ. The LM test set consists of four model families: Falcon (Almazrouei et al., 2023), Qwen (Bai et al., 2023), MPT (Team, 2023), and OLMo (OLMo et al., 2024). Each model family contains four LMs, each intended to test different variations of the shared architectural design in order to study generalized model trends. Importantly, the four individual models in each family are selected such that they can be divided into pairs of the same parameter count (e.g. pairs of 7B/13B sized models for OLMo or 10B/7B for Falcon), and the two models within a pair represent a base (henceforth 'non-instruction tuned') version and an instruction-tuned (either 'Instruct' or 'Chat') version of that model. Therefore, for instance, the two pairs that comprise the Qwen model family are {`(Qwen1.5-7B, Qwen1.5-7B-Chat)`, `(Qwen1.5-14B, Qwen1.5-14B-Chat)`}. A similar logic is applied for model selection among the other model families.

This methodology for model selection allows us to study the effects of intruction-tuning on model accuracy and bias for multiple model sizes, as well as how instruction-tuning generally interacts with the described bias correction methods across different LM architectures.

Evaluation setup

We evaluate LMs using two types of setups: a **zero-shot, no-prompting** setup, where only the test item is presented, and a **few-shot prompting** setup, which provides a one-line task instruction (`Answer the following yes-no questions:`), labels the questions and responses, and includes two examples (one question each using YES and NO answers respectively) created as per the format of the dataset. The motivation here is to test the efficacy of bias correction with varying task/context complexity. While LLMs generally benefit from few-shot prompting (Brown et al., 2020), the higher task demand is shown to be problematic for smaller LMs (Hu & Frank, 2024).

We use LogProbs to derive per-question LM responses, which is calculated as the sum of log probabilities for a certain token, conditioned on the preceding sentence tokens. Specifically, we compare the total LogProbs scores for YES responses (`_Yes` and `Yes`) tokens vs. the total LogProbs scores for NO responses (`_No` and `No`). An aggregated probability for YES and NO is calculated using log-sum-exp. The response category with the highest aggregated LogProbs value is recorded as the model's final decision for the question. Thus, we restrict the space of possible model responses while accounting for some token-level variance. This response derivation method is referred to as 'base inference', and the two described bias correction strategies build on top of this method.



(a) zero-shot prompt

(b) few-shot prompt

Figure 1: Heatmaps showing per-model bias values for all three procedures (Base, Generic, Specific) using zero-shot (a) and few-shot (b) prompts. Positive and negative values indicate yes-bias and no-bias respectively.

Bias Estimation

We define yes-no response bias as:

$$\text{Bias} = \frac{\#\text{YES-responses} - \#\text{NO-responses}}{\#\text{Questions}}$$

On class-balanced datasets (where the expected number of yes and no responses is equal), this metric yields a normalized bias score ranging from -1.0 (pure NO-response behavior, i.e., 'no-bias') to $+1.0$ (pure YES-response behavior, i.e., 'yes-bias'). This metric is consistent with yes-bias estimation in humans (Heather Fritzley & Lee, 2003; Okanda & Itakura, 2008).

Bias Correction Strategies

Method 1: Generic. This correction strategy targets the models' inherent response bias, i.e., the tendency to generally prefer one response over the other, regardless of the context. Such bias may emerge as a function of the pre-training process, where the model learns to place higher probability on tokens that were more prevalent in the training data, dubbed 'common token bias' in (Zhao, Wallace, Feng, Klein, & Singh, 2021). In the context of yes-no bias, a model might be biased towards either YES or NO responses to questions as a result of the prevalence of affirmative or negative response-based text seen during training.

To correct for potential Generic bias, we first calculate 'no-context' YES/NO LogProbs values, i.e., values of YES/NO tokens following the beginning-of-sentence (BOS) token. [2] The no-context scores for both variants are then aggregated across synonymous tokens (e.g., `_Yes` and `Yes`) using log-sum-exp, and subtracted from the LogProbs of YES/NO for each test question. These corrected values are then compared to select the final response.

[2]EOS or PAD is used in case BOS is not available for the tested model.

Method 2: Specific. The model may also exhibit a response bias as a function of the dataset being presented. From a yes-no bias perspective, the model's internal representation of certain subject domains may skew more affirmative or negative to the average, consequently affecting its accuracy. This is highly discussed in the case of controversial political domains such as racial and gender bias (Kaneko, Bollegala, Okazaki, & Baldwin, 2024; Kotek, Dockum, & Sun, 2023), but can also manifest among much more basic, general knowledge questions (such as the ones in our yes-no question datasets).

The Specific yes-no bias correction process involves splitting the question dataset into an 80%-20% train-test set. First, the train set is run through the base inference process. The means of the resulting LogProbs values are used to calculate a bias correction term, c:

$$c = \frac{1}{2 \cdot n_{\text{train}}} \left(\sum_{i=1}^{n_{\text{train}}} \log p(\text{YES} \mid Q_i) - \sum_{i=1}^{n_{\text{train}}} \log p(\text{NO} \mid Q_i) \right)$$

Then, the test set is run through the model, and the bias correction term is applied to the set (added to YES questions and subtracted from NO questions). The splitting and correction is done in a k-fold manner ($k = 5$), and the model response for a question is only recorded when it is part of the test set.

This correction approach operates on the expectation that an unbiased model should yield approximately equal average values for YES and NO on a class-balanced dataset.

Results

Model Families Show Variable Yes-No Bias

Figure 1 shows bias values on COMPS-YNQ and EWoK-YNQ respectively, under both zero-shot no-prompting and few-shot prompting evaluation setups. We find that models do exhibit yes-no bias, and that this bias can be charac-

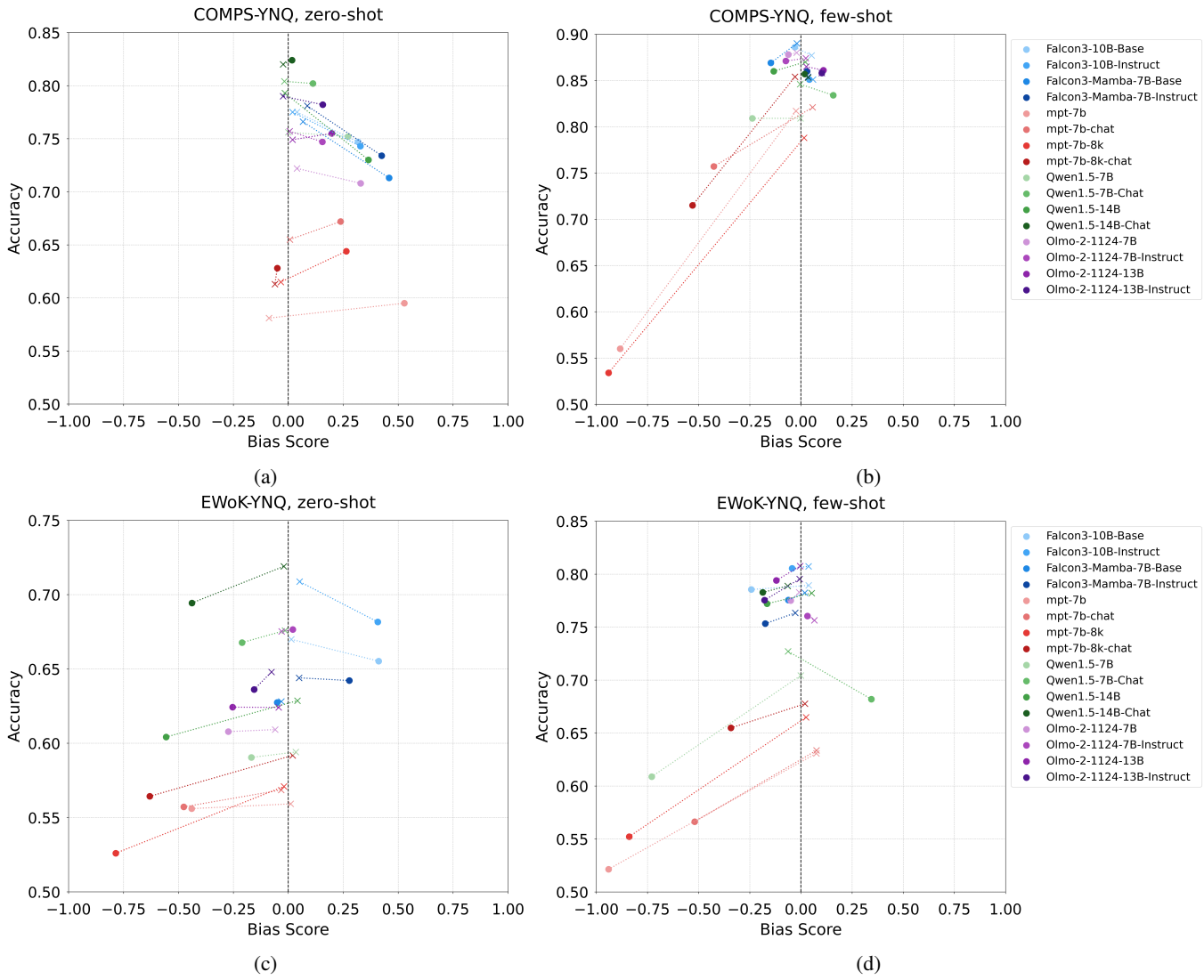


Figure 2: Specific bias correction strategy (×) results in lower bias and similar or higher accuracy relative to Base inference (●) across both datasets (COMPS-YNQ and EWoK-YNQ) and evaluation types (zero-shot and few-shot). Bias values of -1 and +1 indicate pure 'no-bias' 'yes-bias' resp. Bias of 0 indicates unbiased behavior. Accuracy of 0.5 indicates chance performance.

terized at the model family level: all models within a family show preference for the same choice (YES or NO) for a given dataset and prompt setup. However, families as a whole may show different biases depending on the dataset or prompt setup. For an example of dataset-dependent behavior, in zero-shot prompt cases, all model families show some degree of yes-bias on COMPS-YNQ, but three of the model families (MPT, Qwen and OLMo) flip to no-bias on EWoK-YNQ (Fig. 1a base). For examples of prompt-dependent behavior, on COMPS-YNQ (Fig. 1a base), the MPT family exhibits a slight yes-bias with zero-shot prompts, which flips to no-bias with few-shot prompts (Fig. 1b base); the same behavior can be seen in the Falcon family on EWoK-YNQ. In general, looking at base inference bias values, we observe that models lean more towards yes-bias on COMPS-YNQ compared to EWoK-YNQ in a zero-shot setting, but bias behavior in a few-shot setting is similar for the two datasets.

Few-Shot Prompting Reduces Yes-No Bias

A comparison of base inference values between zero-shot and few-shot inputs (i.e. Fig. 1a and 1b base) shows that few-shot prompts result in lower bias values for three model families, with MPT models being a consistent exception. These results align with work from (Kaneko et al., 2024) which showed that few-shot examples can help mitigate gender biases compared to zero-shot prompting in LLMs. However, purely instruction-based debiasing is generally inadequate due to its induced instability, as noted by (Zhao et al., 2021) and (Ma et al., 2023), which explains the MPT family exception here. Additionally, the MPT models contain the lowest number of parameters (6.7B) and training tokens (1T) compared to other '7B' models in our test set (7.3B/1.5T (Falcon3-Mamba),

7.7B/3T (Qwen1.5), and 7.3B/4T (OLMo)), and its behavior here acts as potential supporting evidence for the claim from (Hu & Frank, 2024) that the added task demand of understanding few-shot prompts is detrimental towards the reasoning capabilities of smaller LMs.

Instruction-Tuning Reduces Yes-No Bias

Prompt Type	COMPS-YNQ		EWoK-YNQ	
	Mean Diff.	Cohen's d	Mean Diff.	Cohen's d
zero-shot	-0.156	1.306	-0.040	0.291
few-shot	-0.146	0.837	-0.166	0.736

Table 1: Mean difference in bias between pre-trained and instruction-tuned LM pairs during base inference (without bias correction). Cohen's d ≥ 0.8 indicates a large bias reduction effect size induced by instruction-tuning.

To determine the effects of instruction tuning on model response bias, we compared instruction-tuned LMs and their non-instruct counterparts. The pair-wise analysis in Table 1 shows that, on average, the instruction-tuned versions of LMs are less biased for both datasets and prompt types. All mean differences in bias between the corresponding model pairings are negative, indicating consistent bias reduction in instruction-tuned models. Cohen's d is defined as the number of standard deviations between the mean biases of instruct and non-instruct models, with values higher than 0.8 indicating a strong effect of bias reduction. This result additionally acts as support for work by (Aw, Montariol, AlKhamissi, Schrimpf, & Bosselut, 2024) who show that instruction-tuning can improve world knowledge representations as well as brain alignment in LLMs.

Efficacy of Bias Correction Methods

Prompt Type	Correction Method	COMPS-YNQ		EWoK-YNQ	
		Δ% Bias	Δ% Acc.	Δ% Bias	Δ% Acc.
zero-shot	Generic	-150.90	-13.06	-135.96	-6.09
	Specific	$\mathbf{-101.74}$	$+1.36$	$\mathbf{-90.57}$	$+2.12$
few-shot	Generic	-55.83	$+2.13$	-207.79	-0.86
	Specific	$\mathbf{-93.23}$	$+7.20$	$\mathbf{-106.13}$	$+5.26$

Table 2: Relative percentage difference of mean accuracy and bias for both correction methods compared to base inference values. For bias, a value near -100% indicates complete elimination of bias (highlighted), $< -100\%$ indicates over-correction towards the opposing response, and $> -100\%$ indicates under-correction within the current response category.

Effect on Bias From the two described bias correction strategies, Specific correction emerges as the more reliable method to reduce yes-no bias. Figure 2 compares the model performances between base inference and Specific correction on COMPS-YNQ and EWoK-YNQ. The dotted lines that trace the accuracy and bias changes between the base inference and Specific methods show that the Specific correction consistently brings the model bias close to 0.

On the other hand, Generic correction has mixed effects on model bias. Table 2 reports the model performance resulting from this correction. On applying the Generic correction, we observe cases of both drastic bias over-correction, where the average model bias flips and doubles in magnitude in EWoK-YNQ (few-shot), as well as under-correction where only $\approx 56\%$ of the existing bias was mitigated on average in the case of COMPS-YNQ (few-shot). In comparison, Specific correction is much more consistent and effective, with all percentage differences close to -100% (highlighted in Table 2).

Effect on Accuracy Table 2 shows that the Specific method has a small positive effect on overall model accuracy across both datasets and prompt types. We see more substantial accuracy improvements in individual model families, especially MPT with few-shot prompting (Figs. 2b and 2d) and Falcon with zero-shot prompting (Figs. 2a and 2c). The only exception where we observe an accuracy decrease is for MPT on COMPS-YNQ with zero-shot prompting, although it shows considerable accuracy increases in all other prompt-dataset configurations. Interestingly, this is also the only case where MPT shows a base yes-bias rather than a no-bias, further showing model performance profiles' dependency on the dataset and prompt-type.

The Generic method, in comparison, again shows mixed effects for model accuracy. While family-wise accuracy effects of Generic correction are not reported in the figures, Table 2 shows a small degradation in overall model accuracy with zero-shot prompts (-13% and -6% on COMPS-YNQ and EWoK-YNQ resp.), and even smaller effects with few-shot prompts ($+2\%$ and -0.9% resp.).

Discussion

We present a comprehensive analysis of yes-no bias in language models, revealing trends in bias behavior for popular model families. We provide two yes-no question datasets, COMPS-YNQ and EWoK-YNQ, aimed at evaluating performance on context-free and context-dependent questions. We describe two LogProbs-based bias correction methods, of which the Specific correction presents a reliable method to reduce yes-no bias while preserving accuracy. The Generic correction method, though not as effective, has theoretically justifiable roots, and presents a second example of using LogProbs values to assess and systematically manipulate model response behavior for bias reduction. Based on the efficacy of the Specific correction method, we believe that applying this correction can be beneficial when probing LM internal knowledge or general capabilities. Additionally, it not only furthers de-biasing efforts in LMs, but also efforts in extracting higher reasoning capabilities from smaller models, where previous work (Hu & Frank, 2024) has shown that smaller LMs possess reasoning that is masked due to the added load

of understanding the task itself.

The results of our study hold interesting implications for understanding human response bias, specifically in how bias-driving mechanisms differ between models and humans.

Firstly, we see that LM families exhibit a mixed yes-no bias profile across datasets and prompt-types. Such behavior contrasts with the evidence of yes-bias in humans (Mehrani & Peterson, 2017; Peterson et al., 1999, among others). However, there also exists work to support the idea that LMs somewhat align to human-like linguistic generalization (Hagendorff, Fabi, & Kosinski, 2023; Hu, Mahowald, Lupyan, Ivanova, & Levy, 2024; Jones, Trott, & Bergen, 2024) – one would subsequently expect models to exhibit yes-bias, but this aspect of human response behavior does not transfer over to models. We posit that this is because yes-bias in humans is primarily a non-linguistic phenomenon; it is driven by higher-level factors such as social conformity and desirability. LM learning and reasoning, on the other hand, occurs purely at the linguistic level, using word co-occurrence statistics (Kauf et al., 2023). Although yes-bias manifests in language patterns, its driving mechanisms are primarily non-linguistic in nature. Our base inference results highlight the idea that some human-like response behavior is difficult to model using the quantitative linguistic techniques that constitute LM learning, and these must be addressed before attempting to view LMs as human-like reasoners.

Secondly, the efficacy of the Specific bias correction method also tells us that the existing model bias – regardless of its driving mechanism – is highly correctable by targeting dataset-driven response behavior. We find that models are prone to dataset-specific bias, which in-fact parallels multiple behavioral studies showing that humans exhibit strong, often detrimental domain-specific biases (Almandoz & Tilcsik, 2016; Asplund, Björk, & Magnusson, 2022). Interestingly, cognitive research shows that this human bias can also be mitigated by targeting domain-driven response behavior, such as excluding domain experts in surveys (Asplund et al., 2022), or adopting classical survey techniques such as the Delphi method (Dalkey & Helmer, 1963) where feedback loops between experts and non-experts typically result in more accurate consensus data. This leads us to posit that models are sensitive to domain-specific bias in a similar manner to humans, but as discussed earlier, their underlying causes are likely different, operating on linguistic and non-linguistic levels respectively.

It is possible that even better bias reduction strategies emerge in the future. Currently, the Generic correction method suffers from aggressive over-correction tendencies, and could become viable with the introduction of a weakening effect before application. Additionally, there may be multiple other LogProbs-based bias reduction strategies that can be described using other theoretical justifications, which would warrant testing and could be performed on these datasets. Further analysis of yes-no bias using LMs with more variation in size and architecture, intermediary prompt complexities, and fine-tuning paradigms will help to refine the implicative claims made here, along with strengthening our understanding of why some response behaviors seem to emerge in both humans and models while others do not.

Conclusion

This paper takes a traditionally cognitive science problem – the investigation of yes-bias in question-answering tasks – and transfers it to LMs. In doing so, we (1) demonstrate that LMs show bias when responding to yes-no questions, but that bias is not humanlike, and (2) propose an effective bias correction method that enables a better assessment of underlying LM knowledge, particularly for smaller and non-instruction-tuned LMs.

Our findings highlight yes-bias as an example of bias that, in humans, is likely driven by higher-level factors such as social influences, and is not directly carried over into the distributional properties of the input. In addition, we comprehensively evaluate LMs to show that a characteristic response bias exists in many models, that it is dependent on the model family, dataset, prompt complexity and instruction-tuning, and that it can be corrected in a dataset-specific manner. Thus, despite the fact that yes-no bias in LMs does not follow a humanlike pattern, measuring and correcting for this bias is an important step toward better evaluations of other cognitive properties in LMs.

References

Almandoz, J., & Tilcsik, A. (2016). When experts become liabilities: Domain experts on boards and organizational failure. *The Academy of Management Journal*, *59*(4), 1124–1149. Retrieved 2025-02-03, from http://www.jstor.org/stable/24758185

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., ... Penedo, G. (2023). *The falcon series of open language models.* Retrieved from https://arxiv.org/abs/2311.16867

Asplund, F., Björk, J., & Magnusson, M. (2022). Knowing too much? on bias due to domain-specific knowledge in internal crowdsourcing for explorative ideas. *R&D Management*, *52*(4), 720-734. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/radm.12517 doi: https://doi.org/10.1111/radm.12517

Aw, K. L., Montariol, S., AlKhamissi, B., Schrimpf, M., & Bosselut, A. (2024). *Instruction-tuning aligns llms to the human brain.* Retrieved from https://arxiv.org/abs/2312.00575

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bass, B. M. (1955, November). Authoritarianism or acquiescence? *The Journal of Abnormal and Social Psychology*, *51*(3), 616–623. Retrieved from http://dx.doi.org/10.1037/h0042890 doi: 10.1037/h0042890

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners.* Retrieved from https://arxiv.org/abs/2005.14165

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining* (p. 6437–6447). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3637528.3671458 doi: 10.1145/3637528.3671458

Dalkey, N., & Helmer, O. (1963). An experimental application of the delphi method to the use of experts. *Management Science*, *9*(3), 458–467. Retrieved 2025-02-03, from http://www.jstor.org/stable/2627117

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119-130. Retrieved from https://www.sciencedirect.com/science/article/pii/S0092656615000495 doi: https://doi.org/10.1016/j.jrp.2015.05.004

Fazio, R. H., & Olson, M. A. (2003, February). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327. Retrieved from http://dx.doi.org/10.1146/annurev.psych.54.101601.145225 doi: 10.1146/annurev.psych.54.101601.145225

Fox, W. S., Payne, D. E., Priest, T. B., & Philliber, W. W. (1977, June). Authority position, legitimacy of authority structure, and acquiescence to authority. *Social Forces*, *55*(4), 966. Retrieved from http://dx.doi.org/10.2307/2577566 doi: 10.2307/2577566

Furnham, A. (1986, January). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400. Retrieved from http://dx.doi.org/10.1016/0191-8869(86)90014-0 doi: 10.1016/0191-8869(86)90014-0

Hagendorff, T., Fabi, S., & Kosinski, M. (2023, October). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, *3*(10), 833–838. Retrieved from http://dx.doi.org/10.1038/s43588-023-00527-x doi: 10.1038/s43588-023-00527-x

Heather Fritzley, V., & Lee, K. (2003, September). Do young children always say yes to yes–no questions? a metadevelopmental study of the affirmation bias. *Child Development*, *74*(5), 1297–1313. Retrieved from http://dx.doi.org/10.1111/1467-8624.00608 doi: 10.1111/1467-8624.00608

Hu, J., & Frank, M. C. (2024). *Auxiliary task demands mask the capabilities of smaller language models.* Retrieved from https://arxiv.org/abs/2404.02418

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024, August). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36). Retrieved from http://dx.doi.org/10.1073/pnas.2400917121 doi: 10.1073/pnas.2400917121

Ivanova, A. A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T. H., ... Andreas, J. (2024). *Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models.* Retrieved from https://arxiv.org/abs/2405.09605

Jones, C. R., Trott, S., & Bergen, B. (2024, 06). Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (epitome). *Transactions of the Association for Computational Linguistics*, *12*, 803-819. Retrieved from https://doi.org/10.1162/tacl_a_00674 doi: 10.1162/tacl_a_00674

Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024). *Evaluating gender bias in large language models via chain-of-thought prompting.* Retrieved from https://arxiv.org/abs/2401.15585

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., ... Lenci, A. (2023). Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, *47*(11), e13386. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13386 doi: https://doi.org/10.1111/cogs.13386

Kotek, H., Dockum, R., & Sun, D. Q. (2023). *Gender bias in llms.* Retrieved from https://arxiv.org/abs/2308.14921

Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., ... Wu, B. (2023). Fairness-guided few-shot prompting for large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 43136–43155). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/8678da90126aa58326b2fc0254b33a8c-Paper-Conference.pdf

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010, May). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. Retrieved from http://dx.doi.org/10.1037/a0019216 doi: 10.1037/a0019216

Mehrani, M. B., & Peterson, C. (2017, March). Children's recency tendency: A cross-linguistic study of persian, kurdish and english. *First Language*, *37*(4), 350–367. Retrieved from http://dx.doi.org/10.1177/0142723717694055 doi: 10.1177/0142723717694055

Misra, K., Rayz, J. T., & Ettinger, A. (2023). *Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models.* Retrieved from https://arxiv.org/abs/2210.01963

Okanda, M., & Itakura, S. (2008, March). Children in

asian cultures say yes to yes—no questions: Common and cultural differences between vietnamese and japanese children. *International Journal of Behavioral Development*, *32*(2), 131–136. Retrieved from http://dx.doi.org/10.1177/0165025407087211 doi: 10.1177/0165025407087211

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., ... Hajishirzi, H. (2024). *2 olmo 2 furious.* Retrieved from https://arxiv.org/abs/2501.00656

Paulhus, D. L. (1991). Measurement and control of response bias. In *Measures of personality and social psychological attitudes* (p. 17–59). Elsevier. Retrieved from http://dx.doi.org/10.1016/B978-0-12-590241-0.50006-X doi: 10.1016/b978-0-12-590241-0.50006-x

Peterson, C., Dowden, C., & Tobin, J. (1999, October). Interviewing preschoolers: Comparisons of yes/no and wh- questions. *Law and Human Behavior*, *23*(5), 539–555. Retrieved from http://dx.doi.org/10.1023/A:1022396112719 doi: 10.1023/a:1022396112719

Randall, D. M., & Fernandes, M. F. (1991, November). The social desirability response bias in ethics research. *Journal of Business Ethics*, *10*(11), 805–817. Retrieved from http://dx.doi.org/10.1007/BF00383696 doi: 10.1007/bf00383696

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). *Large language models show human-like social desirability biases in survey responses.* Retrieved from https://arxiv.org/abs/2405.06058

Schoenegger, P., Tuminauskaite, I., Park, P. S., Bastos, R. V. S., & Tetlock, P. E. (2024, November). Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, *10*(45). Retrieved from http://dx.doi.org/10.1126/sciadv.adp1528 doi: 10.1126/sciadv.adp1528

Sinha, K., Gauthier, J., Mueller, A., Misra, K., Fuentes, K., Levy, R., & Williams, A. (2022). *Language model acceptability judgements are not always robust to context.* Retrieved from https://arxiv.org/abs/2212.08979

Steinmetz, J., & Posten, A.-C. (2020). Guidelines on acquiescence in marketing research. *Advances in Consumer Research*, *48*, 712–714. Retrieved from https://www.acrwebsite.org/volumes/2651255/volumes/v47/NA-48

Team, M. N. (2023). *Introducing mpt-7b: A new standard for open-source, commercially usable llms.* Retrieved 2023-05-05, from www.mosaicml.com/blog/mpt-7b (Accessed: 2023-05-05)

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024, 09). Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, *12*, 1011-1026. Retrieved from https://doi.org/10.1162/tacl_a_00685 doi: 10.1162/tacl_a_00685

van de Mortel, T. F. (2008). *The Australian Journal of Advanced Nursing*, *25*(4), 40–48. Retrieved from https://search.informit.org/doi/10.3316/informit.210155003844269

Wright, J. D. (1975). Does acquiescence bias the "index of political efficacy?". *The Public Opinion Quarterly*, *39*(2), 219–226. Retrieved 2025-02-03, from http://www.jstor.org/stable/2748148

Zhao, T., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning.* Retrieved from https://api.semanticscholar.org/CorpusID:231979430

Zou, Z., Mubin, O., Alnajjar, F., & Ali, L. (2024, February). A pilot study of measuring emotional response and perception of llm-generated questionnaire and human-generated questionnaires. *Scientific Reports*, *14*(1). Retrieved from http://dx.doi.org/10.1038/s41598-024-53255-1 doi: 10.1038/s41598-024-53255-1