Estimating and Correcting Yes-No Bias in Language Models

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

Background

Humans are known to show an **acquiescence bias (yes-bias)**:

- Children more often answer 'yes' to unknown questions [1]
- Adults self-report 'yes' more on personality/health/political surveys [2,3]

1. Social Hypothesis:
Yea-saying in humans is a result of social desirability/conformity, aversion of confrontation, other social pressures such as authority.

2. Distributional Hypothesis:
Yea-saying in humans is result of 'yes' appearing more commonly in distributional patterns in our language input.

We can **test** the **distributional hypothesis** by **evaluating** purely statistical learners, i.e., **language models (LMs) for yes bias!**

We find that LMs do not acquiesce like humans, but do exhibit systematic, dataset-dependent yes-no preference. We also present a zero-cost LogProbs-based method to correct "yes-no bias" in LMs.

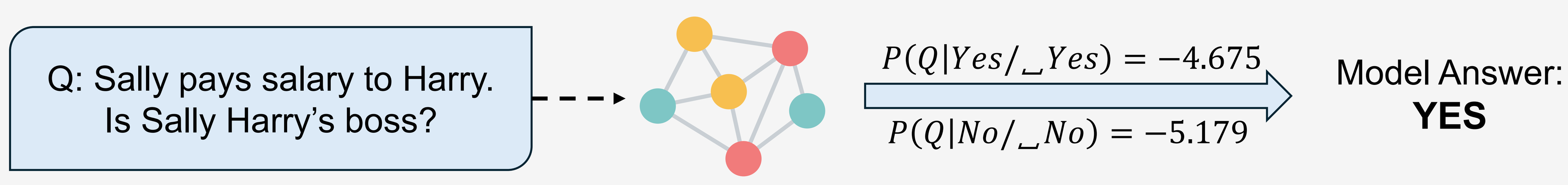
Methods

We measure Yes-No bias as:

$$\text{Bias} = \frac{\#\text{Yes-responses} - \#\text{No-responses}}{\#\text{Questions}}$$

On **class-balanced** datasets, this yields a normalized bias score ranging from -1 (*NO*) to $+1$ (*YES*) response behavior.

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



We explore response behavior and bias correction efficacy in relation to:

- ***Prompt Complexity***: Test LMs using zero-shot & few-shot prompts.
- ***Instruction Tuning***: Test both instruct & non-instruct versions of LMs.
- ***Model Family:*** Test LMs with shared design & training procedures.

Generic Correction

Intuition: LM prefers one response over the other regardless of input.

1. Feed BOS token to LM to get ***'zero-input'*** LogProbs for $Yes/\_Yes$ and $No/\_No$ tokens.
2. Subtract these values from dataset-derived LogProbs values.

Simple baseline strategy, meant to target 'common token' bias [4].

Dataset-Specific Correction

Intuition: LM shows response bias as a function of the dataset being tested.

1. Dataset is split into $\approx 20\%$ ***evaluation sets***, with the remaining $\approx 80\%$ being used as ***calibration sets.***
2. LogProbs for $Yes/\_Yes$ and $No/\_No$ are averaged across the calibration set.
3. Mean values are subtracted from LogProbs values for evaluation set items.

Done in a ***k-fold*** (k=5) fashion for full dataset evaluation.

Datasets

I. COMPS-YNQ: Adapted from COMPS [5], yes-no conversion of 2100 concept-property pairs testing basic world knowledge:

{An iguana/a trolley} basks in the sun. → Does {an iguana/a trolley} bask in the sun?

II. EWoK-YNQ: Adapted from EWoK [6], yes-no conversion of 2056 context-target pairs testing ***contextual*** world knowledge:

Chao is making Yan's job {easier/harder} / Chao is {helping/hindering} Yan → Chao is making Yan's job {easier/harder}. Is Chao {helping/hindering} Yan?

Results

▼ Applying dataset-specific correction *(relative to base inference)*:

Avg. bias reduction % (all models):

Zero-shot:
COMPS: -101.67%
EWoK: -90.57%

Few-shot:
COMPS: -93.23%
EWoK: -106.13%

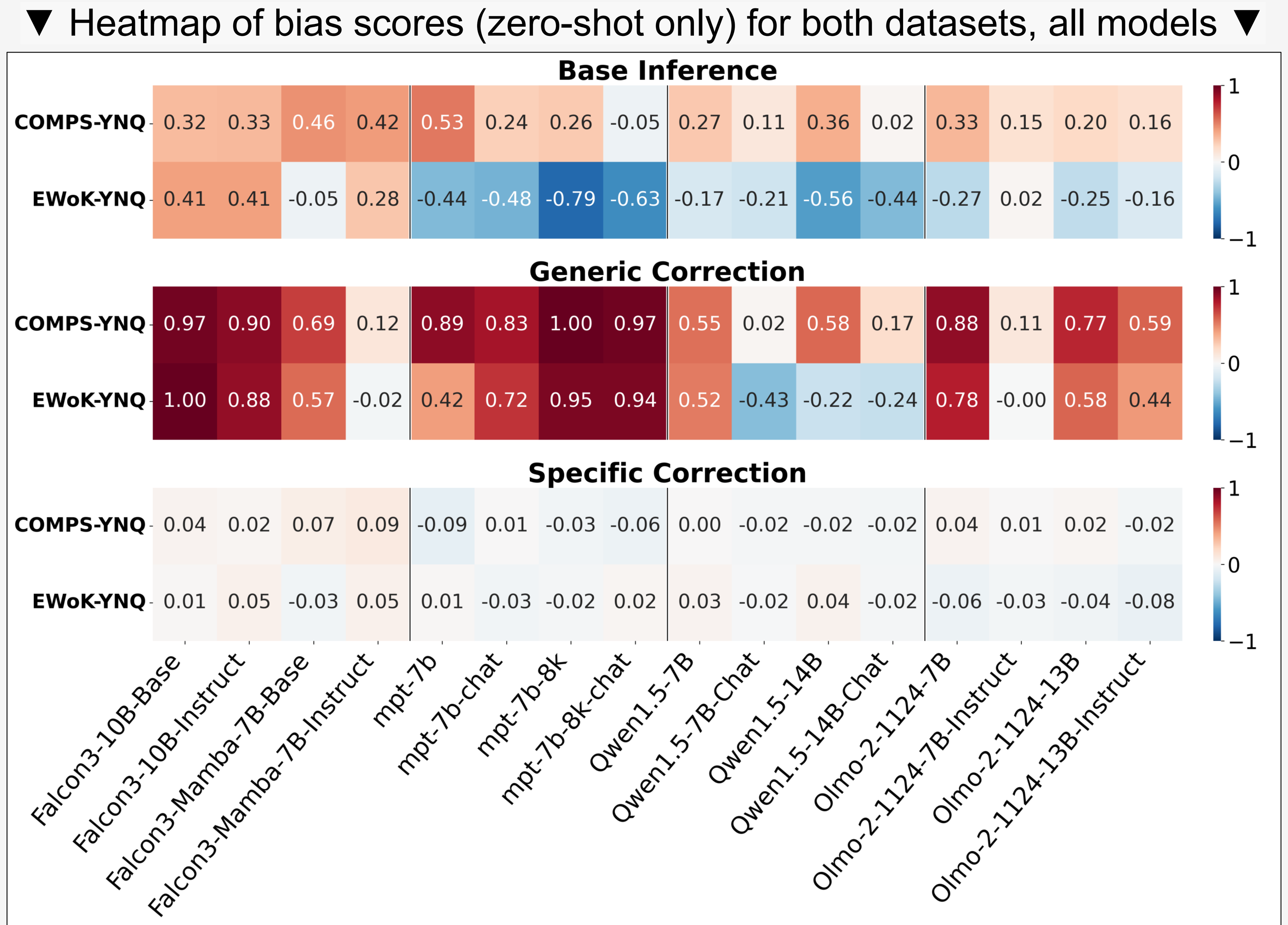
Avg. acc. increase % (all models):

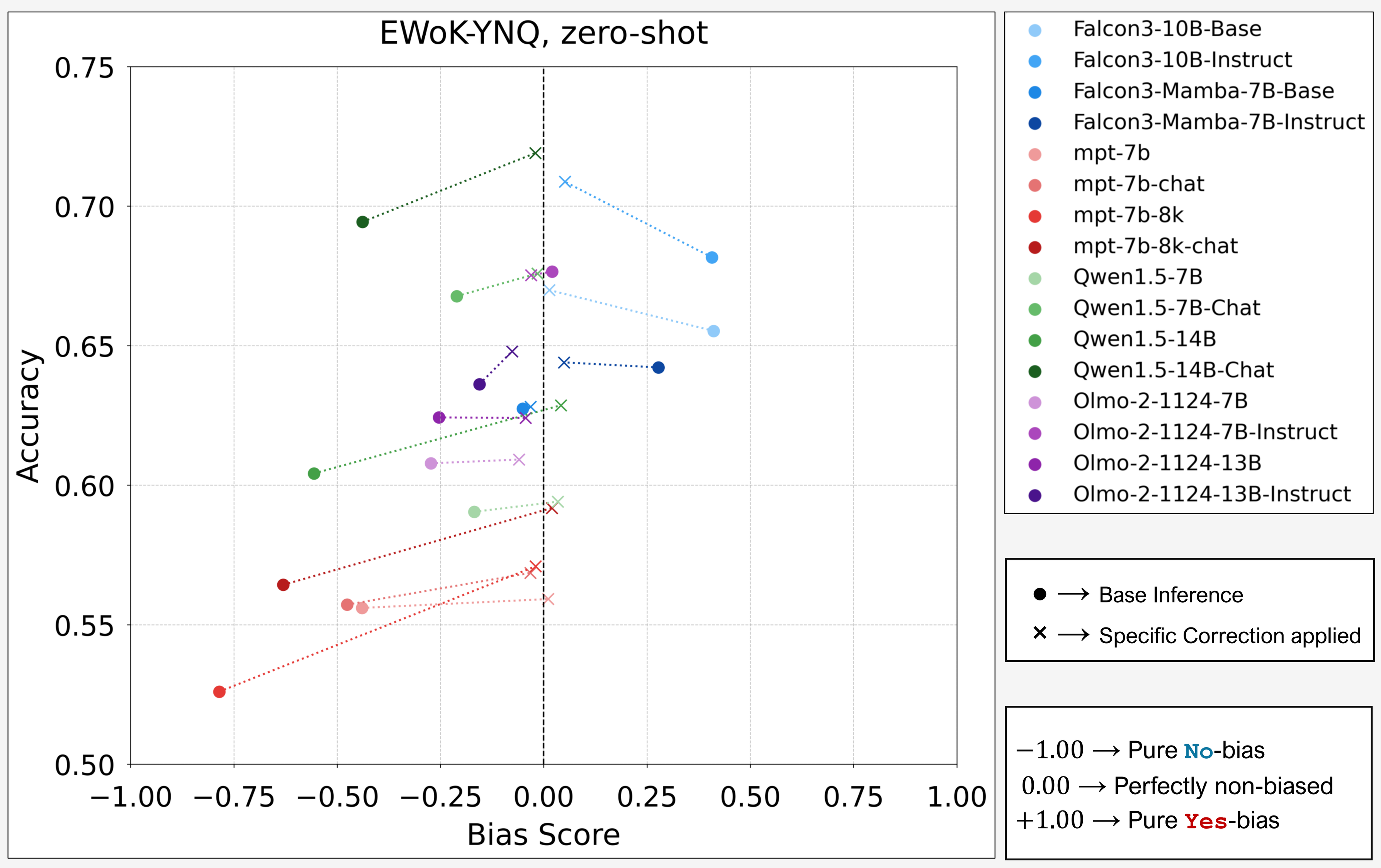
Zero-shot:
COMPS: $+1.36\%$
EWoK: $+2.12\%$

Few-shot:
COMPS: $+7.20\%$
EWoK: $+5.26\%$

- *Generic correction* often *worsens* bias, while **dataset-specific correction drastically improves bias and maintains accuracy!**
- Response preferences are consistent *within* families but vary *across* datasets and prompt complexity.

▼ Heatmap of bias scores (zero-shot only) for both datasets, all models ▼





▲ Well-performing, low-bias LMs should place at the top-center of scatterplot ▲

- **Few-shot** prompting generally has a **bias reduction effect.**
- **Instruction-tuned models** generally show **less-biased behavior** compared to non-instruct counterparts *(~0.155 lower bias score on avg.)*

Future Work

1. We are **adapting** dataset-specific correction **to work on multiple-choice questions:**
   - LMs are known to exhibit **position bias** (primacy/recency effect interfering with content validity) – a similar LogProbs correction strategy could help fix this!
2. We are expanding to **test more datasets** with new subject matter – mathematical ability, logical reasoning, multi-context reasoning, etc.
3. We are testing **smaller-sized calibration sets** – turns out we need *much* lower than 80% to achieve these results!

References

[1] Fritzley, V. H., & Lee, K. (2003). *Do young children always say yes to yes–no questions? A metadevelopmental study of the affirmation bias.* Child Development, 74, 1297-1313. https://doi.org/10.1111/1467-8624.00608

[2] Danner, D., Aichholzer, J., & Rammstedt, B. (2015). *Acquiescence in personality questionnaires: Relevance, domain specificity, and stability.* Journal of Research in Personality, 57, 119-130. https://doi.org/10.1016/j.jrp.2015.05.004

[3] Hill, S. J., & Roberts, M. E. (2023). *Acquiescence bias inflates estimates of conspiratorial beliefs and political misperceptions.* Political Analysis, 31(4), 575-590. https://doi.org/10.1017/pan.2022.28

[4] Zhao, T., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). *Calibrate before use: Improving few-shot performance of language models.* In Proceedings of the 38th International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2102.09690

[5] 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.* In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2210.01963

[6] Ivanova, A., Sathe, A., Lipkin, B., Kumar, U., Radkani, S., Clark, T. H., Kauf, C., Hu, J., Pramod, R. T., Grand, G., Paulun, V., Ryskina, M., Akyürek, E., Wilcox, E., Rashid, N., Choshen, L., Levy, R., Fedorenko, E., Tenenbaum, J., & Andreas, J. (2025). *Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models.* https://doi.org/10.48550/arXiv.2405.09605