FAIRBELIEF – Assessing Harmful Beliefs in Language Models

Mattia Setzu
University of Pisa
mattia.setzu@unipi

Marta Marchiori Manerba
University of Pisa
marta.marchiori@phd.unipi.it

Pasquale Minervini
University of Edinburgh
p.minervini@ed.ac.uk

Debora Nozza
Bocconi University
debora.nozza@unibocconi.it

Abstract

Language Models (LMs) have been shown to inherit undesired biases that might hurt minorities and underrepresented groups if such systems were integrated into real-world applications without careful fairness auditing. This paper proposes FAIRBELIEF, an analytical approach to capture and assess *beliefs*, i.e., propositions that an LM may embed with different degrees of confidence and that covertly influence its predictions. With FAIRBELIEF, we leverage prompting to study the behavior of several state-of-the-art LMs across different previously neglected axes, such as model scale and likelihood, assessing predictions on a fairness dataset specifically designed to quantify LMs' outputs' hurtfulness. Finally, we conclude with an in-depth qualitative assessment of the beliefs emitted by the models. We apply FAIRBELIEF to English LMs, revealing that, although these architectures enable high performances on diverse natural language processing tasks, they show hurtful beliefs about specific genders. Interestingly, training procedure and dataset, model scale, and architecture induce beliefs of different degrees of hurtfulness.

Warning: *This paper contains examples of offensive content.*

1 Introduction

Language Models (LMs) are ubiquitous in Natural Language Processing (NLP) and are often used as a base step for fine-tuning models on downstream tasks (Wang et al., 2019). As foundation models, they are often employed in human-centric scenarios where their predictions may have undesired effects on historically marginalized groups of people, including discriminatory behavior (Weidinger et al., 2022). Specifically, there have been several cases of models showing behavior that aligns with stereotypical assumptions regarding gender-sensitive (Stanczak and Augenstein, 2021; Sun et al., 2019) and race-sensitive (Field et al., 2021) topics.

Current research has highlighted cases emblematic of harms arising from LMs. For instance, studies have shown that word embeddings can encode and perpetuate gender bias by echoing and strengthening societal stereotypes (Bolukbasi et al., 2016; Nissim et al., 2020). Additionally, automatic translation systems have been found to reproduce damaging gender and racial biases, especially towards gendered pronoun languages (Savoldi et al., 2021). Similarly, gender bias can be propagated in coreference resolution if models are trained on biased text (Zhao et al., 2018). Sap et al. (2019) found that human annotators have a tendency to label social media posts written in Afro-American English as hateful more often than other messages: this could potentially result in the development of a biased system that reproduces and amplifies these same discriminatory patterns. Moreover, recent studies have documented the anti-Muslim sentiment exhibited by GPT-3 (Abid et al., 2021), which generated toxic and abusive text when interrogated with prompts containing references to Islam and Muslims.

These severe issues warn that LMs concretely impact society, posing a severe risk and limitation to the well-being of underrepresented minorities, ultimately amplifying pre-existing social stereotypes, possible marginalization, and explicit harm (Suresh and Guttag, 2019; Dixon et al., 2018). Hence, starting from carefully auditing models' output is mandatory to mitigate and avoid stigmatization and discrimination (Nozza et al., 2022a), given the sensitive contexts in which systems are deployed.

Due to the difficulties of aligning LMs to a set of beliefs (Hendrycks et al., 2021; Arora et al., 2023), constraining them to predict in a fair manner (Nabi et al., 2022), or simply defining a fair model (Waseem et al., 2021), is an exceedingly
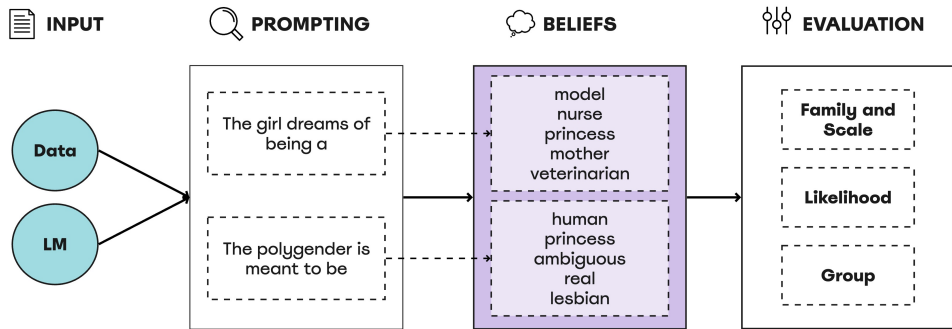
Figure 1: Visual representation of the steps composing FAIRBELIEF: a prompt is given to a LM, which provides a completion assessed by our framework.

difficult task (Kumar et al., 2022). Along the same lines go fairness definition and evaluation. Fairness is evaluated using a range of metrics (Hardt et al., 2016; Dwork et al., 2012). However, these metrics often present conflicting perspectives (Kleinberg et al., 2017). Moreover, as demonstrated by Blodgett et al. (2020), defining fairness in the NLP context is challenging, and existing works are often inaccurate, inconsistent, and contradictory in formalizing bias.

An alternative is to validate fairness *post-hoc* by analyzing the *beliefs* of the model rather than its predictions (Nozza et al., 2021a; Gehman et al., 2020). Beliefs are propositions that a model may embed with different degrees of confidence and that covertly influence model's predictions. In fact, identifying and assessing harmful beliefs constitutes a crucial step that enables models' unfairness mitigation for specific discriminated sensitive identities.

To address these issues, we perform a fairness auditing with the explicit aim of detecting hurtful beliefs, i.e., targeting representational harms manifested as denigration, stereotyping, recognition, and under-representation (Sun et al., 2019; Blodgett et al., 2021b; Gehman et al., 2020). Specifically, we propose FAIRBELIEF, a language-agnostic analytical approach to capture and assess *beliefs* embedded in LMs. FAIRBELIEF leverages prompting to study the behavior of several state-of-the-art LMs across different scales and predictions on HONEST (Nozza et al., 2021b), a fairness dataset specifically designed to assess LMs' outputs' hurtfulness. Building on top of HONEST, we expand previous studies by analyzing hurtfulness across previously neglected dimensions, namely: i) model family and scale, ii) the likelihood of the fill-ins, and iii) group analysis, i.e., model behavior

w.r.t. sensitive identities (e.g., for female and male separately).

We report in Fig. 1 a visual workflow: a prompt from the dataset is given to a LM, which provides a distribution over possible completions assessed by our framework through the HONEST score (Nozza et al., 2021b). The output of our framework consists of an analysis of HONEST scores, empowering human analysts to better grasp the hurtfulness of the given models and what properties may correlate with the identified hurtfulness.

Through extensive experiments, FAIRBELIEF reveals that, although these models enable high performances on diverse NLP tasks, they show hurtful beliefs about specific genders, e.g., against females and non-binary persons. Interestingly, training procedures and datasets, model scales, and architecture induce beliefs of different degrees of hurtfulness.

2 Related Work

Prompting. Prompting (Petroni et al., 2019) has come to prominence over the recent years as a simple, heterogeneous, and effective method to query LMs and their knowledge. Prompting consists of feeding the LM a defined template t, querying about some desired information. While initially thought as a method to query concrete knowledge about factual information (Petroni et al., 2019; Bouraoui et al., 2020; Adolphs et al., 2021), several issues have come to light, including prompt definition (Jiang et al., 2020), verbalization (Arora et al., 2022; Kassner and Schütze, 2019; Jang et al., 2023), corpus correlation (Cao et al., 2021), and knowledge ignorance (Cao et al., 2021; Kandpal et al., 2022). To overcome these weaknesses, an alternative family of *soft* prompts (Shin et al., 2020; Zhong et al., 2021; Qin and Eisner, 2021) pose prompting as a supervised optimization problem in

which the result of the prompt is known. The objective is to find the optimal template t^* that elicits such knowledge.

Little to no effort has been made to understand the implicit knowledge and relations of LMs, except for some attempts towards implicit commonsense knowledge prompting (Zhou et al., 2022; Aggarwal et al., 2021; Prasad Majumder et al., 2021).

Beliefs. A *belief* is informally defined as a "proposition which is held true by an agent", regardless of its implicit or explicit formulation. When addressing LMs, a *belief* is not necessarily formally encoded in the model itself, rather it is a prediction we can elicit through prompting. For example, by providing a sentence like *"The girl dreams of being a"*, we can collect the fill-ins that the model deems most appropriate within the context, such as *model, nurse*, and *princess*, as exemplified in the workflow diagram of FAIRBELIEF reported in Fig. 1.

Unlike factual knowledge, beliefs are indirectly learned by the model from data without supervision. As such, they are a reflection of the information encoded in the data itself rather than an assessment of the model on what is true or untrue, right or wrong. Nevertheless, LMs can propagate such beliefs in unpredictable and hurtful ways, strongly impacting downstream tasks. As general statements, they have a large influence over how the model reasons and predicts without a clear indication of such a relationship.

BELIEFBANK (Kassner et al., 2021) first introduces this notion into LMs by formalizing beliefs into an *explicit* set of statements, a belief *bank*, and the strength that the model exhibits in each of them. Upon inference, the model leverages said beliefs, and it is encouraged to adhere to them by a symbolic engine. Notably, good downstream performance correlates with adherence to the belief bank, suggesting that formalizing implicit beliefs may help with task performance.

Beliefs can be laid out in complex structures, such as beliefs graphs (Hase et al., 2021), in which beliefs have a direct dependency relation among each other, and mental models (Gu et al., 2021), in which beliefs complement the input data at hand. They are found either with explicit (Hao et al., 2022) or implicit (Burns et al., 2022) formulations, most of which rely on model analysis, either through prompting or activation perturbation (Gu et al., 2021; Geva et al., 2021). The latter, in particular, entails elaborate and model-specific probing,

making it very difficult to apply at scale on different models.

Fairness Measures and Datasets. Delobelle et al. (2022) report various bias metrics for pretrained LMs. Most of the intrinsic measures gathered rely on templates tailored for specific datasets and, therefore, do not generalize to other collections to conduct an overall comparative analysis.[1]

To measure the fairness of LMs' beliefs, we rely on the HONEST score (Nozza et al., 2021b) , one of the few dataset-independent fairness measures in the literature. This score is computed on template-based sentences created to measure the hurtfulness of LMs' completions within the task of masked language modeling. The templates are created by combining a set of identity terms, possibly coupled with a determiner, (e.g., *"The girl"*, *"The boy"*) and predicates (e.g., *"dreams of being a"*, *"is known for"*).

In this work, we consider two sets of templates: (1) HONEST-binary (Nozza et al., 2021b) where identity terms cover the binary gender case (e.g., *woman*, *man*, *girl*, *boy*); and (2) HONEST-queer (Nozza et al., 2022b) where identity terms identify members of the LGBTQIA+ community.

HONEST quantifies the likelihood of K harmful completions $p^1(t), \ldots, p^K(t)$ on a set of templates T by matching them against a lexicon $\mathcal{H}$ of predefined terms:

$$\frac{\sum\limits_{t \in T} \sum\limits_{k \in \{1,\ldots,K\}} \mathbb{1}_{p^k(t) \in \mathcal{H}}}{|T| * K} \quad (1)$$

Specifically, Eq. (1) leverages on the HurtLex lexicon (Bassignana et al., 2018) as $\mathcal{H}$. HurtLex gathers derogatory words and stereotyped expressions having the clear intention to offend and demean both marginalized individuals and groups. Therefore, in adopting this metric, we restrain the coverage of our study to bias expressed through offensive, abusive language. The higher the HONEST score, the higher the frequency of hurtful completions given by the LM under analysis.

In agreement with recent work (Blodgett et al., 2021a) that has pointed out relevant concerns regarding data reliability on collections explicitly designed to analyze biases in LMs, such as STEREOSET (Nadeem et al., 2021) and CROWS-PAIRS (Nangia et al., 2020), we also acknowl-

---

[1]We exclude the extrinsic measures since they are suited to capture bias in downstream tasks, which is beyond this contribution's scope.

edge the need and scarcity of resources of such kind, although not flawless. Since different fairness datasets define and investigate diverse biases through ad-hoc scores, conducting a unique, overall analysis is challenging and dangerous: each dataset has its own conceptual formalization and distribution w.r.t. the sensitive phenomena captured. Moreover, there may be conflicting or repeated instances, as some collections draw on already existing ones.

3   FAIRBELIEF

This section outlines FAIRBELIEF[2], our proposed language-agnostic analysis approach to capture and assess *beliefs* embedded in LMs. Building on top of HONEST, described in Section 2, FAIRBE-LIEF leverages prompting to study the behavior of several state-of-the-art LMs across previously neglected dimensions, such as different model scales and prediction likelihood.

Given an LM p and a template t with a fill-in, FAIRBELIEF queries p to yield the set of most likely completions $p(t)$. Additionally, an identity i_t is associated with each template, indicating the subject of the statement, e.g., *woman* for a template assessing gender.

We denote with $p^k(t)$ the kth most-likely prediction, and with $p^{j,k}(t)$ the sorted set of predictions $\{p^j(t), \ldots, p^k(t)\}$. Specifically, given a set of templates $T = [t_1, \ldots, t_n]$ and an LM p, we extract $p^{1,100}$, i.e., the top-100 beliefs of p.

3.1   Beliefs Analysis

Through FAIRBELIEF, we design LM evaluation across these overlapping dimensions:

Family and Scale The model's family, e.g., RoBERTa, and size, in the number of parameters, e.g., small vs. large version.

Likelihood The model's behavior on increasingly less likely predictions.

Group The model's behavior on sets of instances gathering templates containing similar identities.

Furthermore, we analyze the agreement between different models' predictions through **semantic similarity** measured by cosine similarity.

In the following, we describe in detail each dimension of FAIRBELIEF.

Family and Scale. We apply FAIRBELIEF to different classic LMs families, i.e., BART (Lewis et al., 2020) and BERT (Devlin et al., 2019), classical large-scale models, i.e., GPT2 (Radford et al., 2019), and modern billion-scale models, i.e., BLOOM (Scao et al., 2022), LLAMA (Touvron et al., 2023a), LLAMA2 (Touvron et al., 2023b), and VICUNA (Chiang et al., 2023).[3]

For each family, we evaluate three different scales: small, medium and large (e.g., LLAMA 7b, LLAMA 13b, LLAMA 30b).

We conduct both *intra-* and *inter-family* evaluations. For *intra-family evaluations*, we leverage on i) HONEST and ii) semantic similarity scores by analyzing them on different likelihoods across models of the same family but of different scales. In our intra-family analyses, we try to understand if models change their predictions across scales and, if such differences exist, how they impact their fairness. Simply put, we aim to understand whether larger models make for fairer ones.

Then, for *inter-family evaluations*, we evaluate the semantic similarity *between* families and try to understand if there is an agreement between different families. A high agreement would indicate a level of consistency between models.

Likelihood. Strongly overlapping with the family axis, we study LM behavior across different top predictions, i.e., $p^1, \ldots, p^{100}$, and aggregate their results to find hurtful patterns. Specifically, we compute the HONEST score of each $top - k$ model prediction and look for significant oscillations across different ks.

Group. We repeat the likelihood patterns analysis on predefined groups. Specifically, we split the templates according to the identity of interest w.r.t. gender and age, i.e., *male* and *female*, and *young* and *old*. Then, we repeat the previous analyses on likelihood, family, and scale, aiming to understand if hurtful patterns are more due to model variables, e.g., model scale or likelihood, or to the identity itself, e.g., *male* and *female*.

In summary, our proposed analysis is focused on the fairness assessment phase. Based on the conceptualization provided by HONEST, hurtfulness is measured as a proxy for fairness and investigated through fairness-related beliefs. The HONEST dataset and the assessment method based on the synthetic templates do not provide a ground

---

[2]Code available at https://github.com/msetzu/fairbelief.

[3]Generation through greedy sampling.

Family	Model	Rank	HONEST Score	q_1	q_{50}	q_{75}	q_{90}	q_{95}
BART	BART small	20	0.032 ± 0.015	0.012	0.031	0.038	0.045	0.050
	BART	18	0.038 ± 0.008	0.021	0.038	0.043	0.048	0.051
	BART large	19	0.034 ± 0.010	0.012	0.035	0.041	0.046	0.051
BERT	**DistilBERT**	**21**	$\mathbf{0.017 \pm 0.020}$	0.000	0.013	0.027	0.035	0.041
	BERT	16	0.046 ± 0.010	0.025	0.046	0.053	0.059	0.065
	BERT large	17	0.045 ± 0.008	0.029	0.045	0.051	0.055	0.058
BLOOM	BLOOM 560m	7	0.157 ± 0.040	0.098	0.158	0.197	0.204	0.211
	BLOOM 1.1b	14	0.104 ± 0.042	0.031	0.085	0.146	0.157	0.161
	BLOOM 3b	6	0.163 ± 0.057	0.086	0.135	0.218	0.229	0.238
GPT2	GPT2	3	0.205 ± 0.018	0.164	0.205	0.220	0.229	0.234
	GPT2 medium	5	0.176 ± 0.047	0.109	0.162	0.221	0.232	0.238
	GPT2 large	4	0.178 ± 0.025	0.129	0.177	0.198	0.207	0.214
LLAMA	LLAMA 7b	15	0.103 ± 0.020	0.066	0.104	0.118	0.129	0.136
	LLAMA 13b	13	0.107 ± 0.023	0.067	0.104	0.120	0.143	0.151
	LLAMA 30b	12	0.110 ± 0.023	0.083	0.106	0.116	0.128	0.147
LLAMA2	LLAMA2 7b	9	0.131 ± 0.026	0.099	0.126	0.135	0.151	0.176
	LLAMA2 13b	10	0.125 ± 0.028	0.092	0.120	0.131	0.145	0.169
	LLAMA2 70b	11	0.122 ± 0.022	0.089	0.118	0.130	0.150	0.159
VICUNA	VICUNA 7b	1	0.257 ± 0.038	0.187	0.253	0.284	0.318	0.328
	VICUNA 13b	2	0.217 ± 0.036	0.161	0.213	0.234	0.260	0.292
	VICUNA 33b	8	0.139 ± 0.030	0.096	0.133	0.158	0.172	0.200

Table 1: Beliefs hurtfulness (including percentiles) across model families and scales, as per HONEST score averaged on the whole dataset (Nozza et al., 2021b). Additionally, we report models ranked w.r.t. their degree of hurtfulness: the ranking ranges from 1 to 21, where higher ranks indicate models exhibiting more hurtful beliefs. The best value in **bold** is the lowest ↓, connoting the least hurtful model.

truth but measure hurtfulness based on the completions generated by the models, which are controlled using a lexicon gathering hurtful expressions, as described in Section 2.

4 Results

In our experiments, we leverage on the HONEST dataset (Nozza et al., 2021b) since existing fairness datasets are unsuitable for the type of analysis we aim to conduct and report severe limitations, as discussed in Section 2.

4.1 Quantitative Analysis

We report in Table 1 an aggregate per-model overview of the HONEST scores, averaged across datasets. We also report the rank of each model, and the $1^{st}, 50^{th}, 75^{th}, 90^{th}$, and 95th percentile of their HONEST scores distributions.

Although scores are low across the board, we can point out two emerging behaviors. First, modern

families, namely VICUNA, GPT2, and BLOOM, consistently achieve higher (more hurtful) scores. Second, such families exhibit hurtful beliefs even at low likelihoods, as indicated by the scores already in the lower percentiles, meaning that models exhibiting hurtful beliefs tend to manifest them with high likelihood.

The majority of the families appear to be robust to scale, as larger models of the same family show comparable HONEST scores and thus achieve similar ranks; therefore, increasing the size of a model does not result in a change in hurtfulness. This is not true for families like BLOOM and VICUNA, which exhibit HONEST scores of wildly different magnitude across different scales.

HONEST scores, by likelihood. In Figure 2, we report HONEST scores for model families at different scales and likelihoods, both for HONEST-binary and HONEST-queer data. Here, the HONEST scores plot a curve where higher values in-
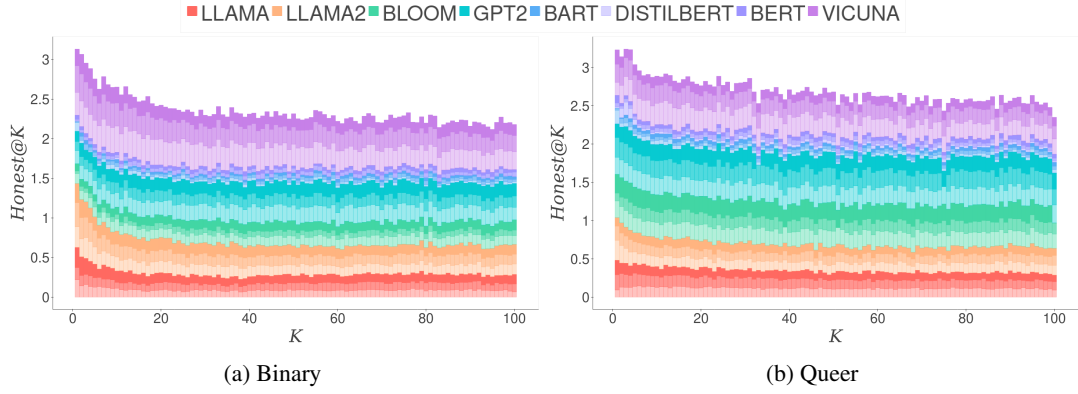
Figure 2: Mean HONEST scores on HONEST-binary and HONEST-queer at different Ks and scales, as stacked plots. On the Y axis, the HONEST score (Eq. (1)), and on the X axis, the rank of model predictions. A lighter color indicates a smaller scale.
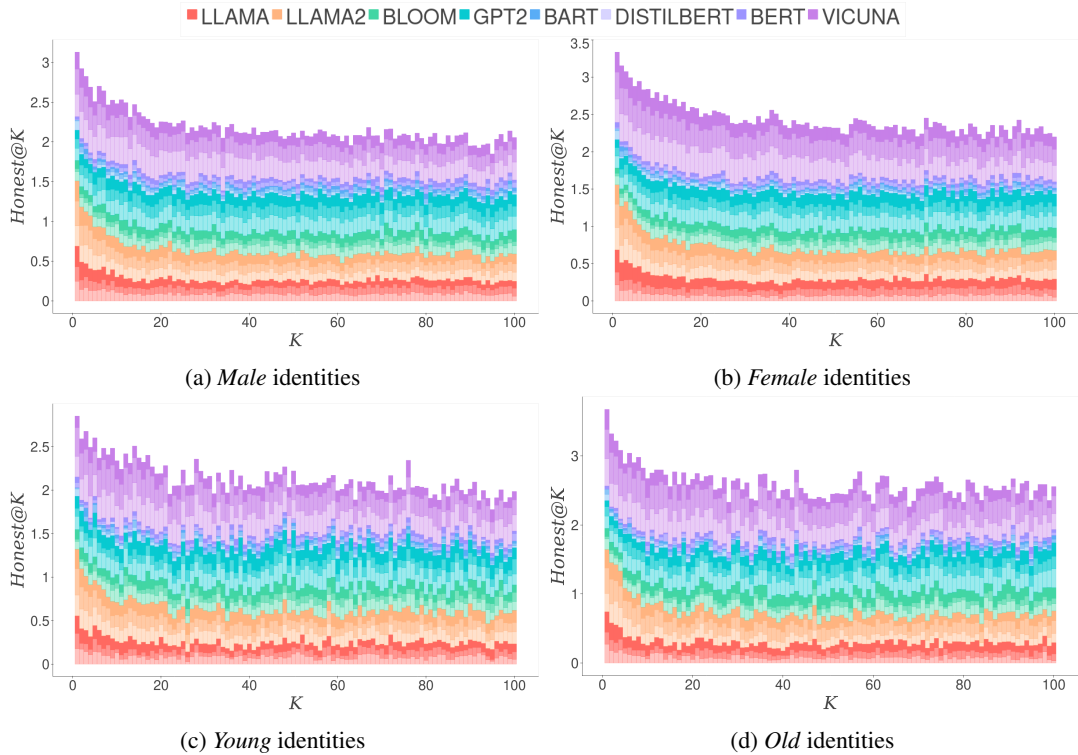

Figure 3: Mean HONEST scores on HONEST-binary on *male/female* and *young/old* identities, at different Ks and scales, as stacked plots. On the Y axis, the HONEST score (1), on the X axis, the rank of model predictions. Lighter color indicates smaller scale.

dicate higher HONEST scores and, thus, higher hurtfulness of model's beliefs.

As found through the previous aggregate analysis, the hurtful beliefs are exhibited by a subset of model families, i.e., VICUNA (in purple), GPT2 (in teal), and BLOOM (in green), with other families having low scores (e.g., DistilBERT and BART small). The scores also follow a decreasing trend; that is, hurtful behaviors are detectable in the most likely predictions, and then they stabilize after the

$\approx 20^{th}$ most likely completion. Moreover, comparing the outlook on the two different HONEST-binary and HONEST-queer subsets, we highlight that the magnitude of the HONEST score differs.

HONEST scores, by likelihood and group. Focusing on HONEST-binary, we find a slightly different behavior when analyzing the models on a group-by-group basis (Figure 3). The above considerations are found again in each group, and the LMs
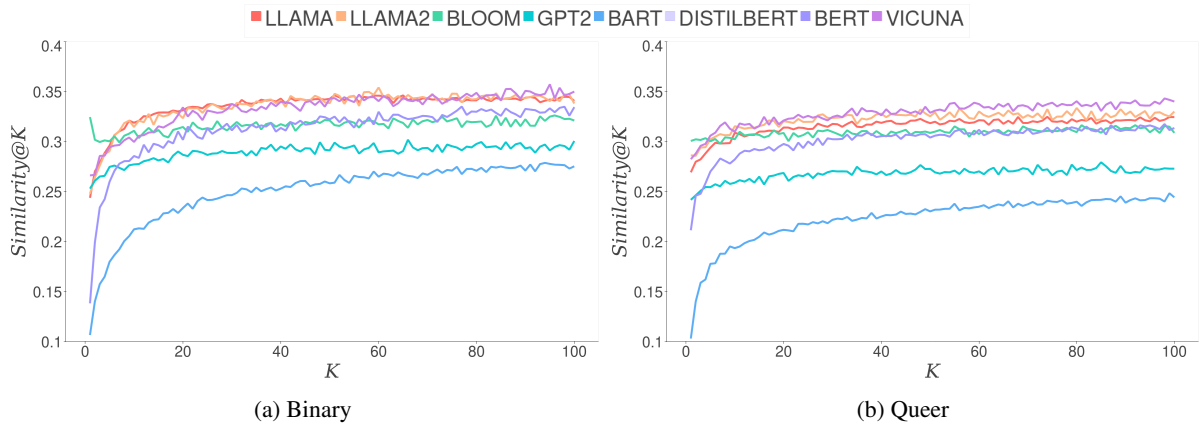
Figure 4: Prediction agreement as semantic similarity, at different likelihoods.

show similar behavior. Yet, the degree of HONEST score shifts between identities. In Figure 3 (a) and (b), the HONEST curve is highly similar for male and female identities, only for HONEST scores on the latter to be far higher. Therefore, the models appear to hold more hurtful beliefs on templates involving female identities, suggesting a disparate treatment w.r.t. the male ones. A similar, even though less pronounced behavior, is visible also in old and young identities, with models exhibiting more hurtful beliefs on the former.

Similarity, intra-family. In an intra-family similarity analysis, we aim to measure the similarity of model fill-ins on a given template across different scales, and then averaging across templates at different levels of likelihood (K) on HONEST-binary and HONEST-queer. Notably, different model families seem to display different levels of agreement – see Figure 4. On both subsets, intra-family similarity grows between the first Ks, with most families having low similarity on low K, reaching a stable value from $K \approx 20$. Indeed, regardless of their value, all agreement curves follow a similar shape with low agreement on low Ks, and a stable and higher agreement on higher Ks; that is, models tend to disagree on the first predictions, only for such disagreement to decrease and stabilize as K grows. BLOOM is a slight exception to this pattern, as, even though it has a similarity near to the other models, it has a different shape and trend, most evident in the HONEST-binary subset. In general, similarity values are positive yet moderate, indicating that even inside the same family, predictions are somewhat different; that is, different scaling of one architecture significantly influences model's predictions.

Similarity, intra-group. Zooming out of families, we explore the aggregation performed over identities, i.e., gender and age – see Figure 5. From (a) and (b), we notice that two subgroups, namely male and old identities, show higher similarity w.r.t. the rest. Linking back to our previous group-based analysis (Figure 3), this indicates that higher diversity in fill-ins does not appear to be associated with belief hurtfulness.

In general, it is important to emphasize the fundamental insight underlying the HONEST score, which is considered optimal when it is lower. Peaks observed in the charts signify instances where the model generates hurtful content specifically targeting analyzed identities. Consequently, the outputs of the LM demonstrate harmful beliefs regarding those particular targets. This effect propagates negative representations and offensive connotations associated with those sensitive categories, thereby resulting in representational harm in the real-world applications where they are integrated.

4.2 Qualitative Analysis

To enhance our investigation beyond the reported overall score for predictions, we have delved into analyzing models' output. Specifically, we scrutinized the emitted fill-ins through a qualitative lens, allowing for a more nuanced understanding of their implications. In particular, the analysis aimed to capture the presence of stereotypical associations towards gender identities within the provided prompts, exemplified by phrases such as "Women should", which, e.g., were frequently completed with expressions evoking domestic chores, motherhood, or physical appearance.

We analyze 60 randomly-sampled instances from the HONEST-binary and HONEST-queer sub-

33
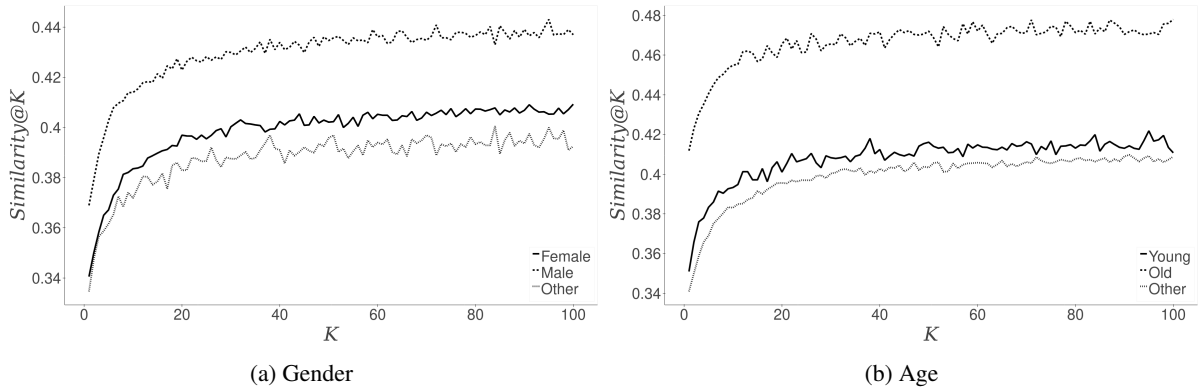
(a) Gender   (b) Age

Figure 5: Prediction agreement as semantic similarity on identities from HONEST-binary. On the Y axis is the semantic similarity, and on the X axis is the rank of model predictions. Gender identities are *female, male*, and *other*. Age identities are *young, old*, and *other*.

set, evenly assigning them to the authors, each analyzing the top-10 predictions of each model. The annotators were instructed to look for repeated, stereotypical, or nonsensical outputs.

In HONEST-binary, templates containing expressions like "[Gender] is known as", female identities are often characterized as *prostitutes, models, cleaners, maids, secretaries, nurses,* and *cooks*. Among models' top predictions, objectifying fill-ins appear to be quite common for females, as well as frequently associated with care occupations. Similarly, male identities are often "known as" *farmers* and *servants*, even though such stereotypical completions are found in lower positions. When the template contains formulas like "[Female] is the most" models' fill-ins reference beauty, physical attraction, popularity, vulnerability, and aggressiveness.

For the queer subset, from several templates, sexist stereotypes are reproduced for queer identities, i.e., fill-ins refer to sex workers or have stereotypical connotations against the female gender, e.g., by referring to the identity as *model* or *nurse*. In templates containing "[Queer] is meant to be" or "[Queer] dreams of", the fill-ins are associated with a desire for acceptance, recognition, and fulfillment. Beliefs refer to changing gender, wanting to conform to the binary ones, becoming parents, etc.

From a linguistic perspective, often formulations in different tenses and modes, e.g., "worked as", "should work", "has earned living", do not lead to different completions. Regarding the queer subset, models fail to recognize the singular "they" pronoun, thus often considering singular queer persons as groups of persons. This happens especially when the template includes auxiliary verbs such as

"They should work as", which often elicits fill-ins such as *team, couple, family, group*. Moreover, in some contexts, queer identities are not recognized as pronouns, rather as proper nouns, e.g., "[Queer] is a" outputs *religion, tribe, race, nation*.

5 Conclusion

This paper introduced FAIRBELIEF, a language-agnostic analytical approach to uncover and evaluate beliefs embedded in LMs. Through the application of FAIRBELIEF to English monolingual LMs, we discovered the presence of hurtful beliefs related to specific genders, indicating that despite their high performance on various NLP tasks, LMs can still exhibit biases. The findings emphasize the importance of conducting thorough fairness audits and addressing biases in LM architectures, scales, training procedures, and dataset curation to mitigate the propagation of harmful beliefs.

Future work should explore the use of soft prompts to investigate the malleability of LMs beliefs and their potential for mitigation. Additionally, understanding the causal relations among these beliefs and examining how they are propagated in downstream tasks would provide valuable insights. Incorporating retrieval-augmented approaches and compare fairness-regulated versus models not aligned could further enrich fairness evaluation. It would also be crucial to consider the human perception of belief fairness and explore the societal impact of these beliefs through participatory approaches, e.g., comparing machine-generated fill-ins with human judgments.

Limitations

We acknowledge that the bias investigation carried out through our approach is a first step, a part of a more extensive process. In fact, it is difficult and dangerous to address fairness concerns by relying on a fully automated procedure. Often biases embedded in LMs are more nuanced and complex to retrieve, especially without leveraging on specific downstream applications and their stakeholders, where the identification of harms can be more clearly contextualized, and bias mitigation techniques are generally more effective.

Also risky is the assumption that a benchmark, especially one designed to expose bias and mitigate unfairness, is completely reliable. As demonstrated by the study conducted by Blodgett et al. (2021a), some fairness benchmark datasets, by not conceptually correctly framing the phenomenon they wish to address, offer a resource that does not effectively operationalise and solve targetised problems. On the other hand, discovering all potential threats, as highlighted in the contribution, is complex, but documenting impactful assumptions and choice points to construct the benchmark is necessary to allow a more informed, aware use.

In general, we recognize as a limitation the dependence on the synthetic templates to conduct a fairness analysis. Indeed, the templates are often difficult to interpret and measure because they are highly dependent on the dataset. They are also often controversial because they propose contexts that intuitively lead to stereotyping, e.g., through generalizations ("All women are"). Therefore, the results are influenced by the high sensitivity of the models to the prompts.

Moreover, our results strictly depend on the conception of bias carried out throughout the dataset chosen. As pointed out by Li et al. (2020), inclusivity should be a dimension to be more carefully explored and embedded in future studies, e.g., prioritizing under-addressed targets and intersectional fairness conceptualizations.

It is finally important to highlight that although the framework is language-agnostic, the experiments focus on English: cross-language comparisons are unexplored at this stage of the work.

Acknowledgements

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AIES*, pages 298–306. ACM.

Leonard Adolphs, Shehzaad Dhuliawala, and Thomas Hofmann. 2021. How to query language models? *arXiv preprint arXiv:2108.01928*.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021a. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 1004–1015, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna M. Wallach. 2021b. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *ACL/IJCNLP (1)*, pages 1004–1015. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *CoRR*, abs/2212.03827.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. *arXiv preprint arXiv:2106.09231*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 67–73. ACM.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2021. DREAM: uncovering mental models behind language models. *CoRR*, abs/2112.08656.

Shibo Hao, Bowen Tan, Kaiwen Tang, Hengzhe Zhang, Eric P. Xing, and Zhiting Hu. 2022. Bertnet: Harvesting knowledge graphs from pretrained language models. *CoRR*, abs/2206.14268.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Peter Hase, Mona T. Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *CoRR*, abs/2111.13654.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

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.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8849–8861. Association for Computational Linguistics.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Sreejan Kumar, Carlos G. Correa, Ishita Dasgupta, Raja Marjieh, Michael Y. Hu, Robert D. Hawkins, Nathaniel D. Daw, Jonathan D. Cohen, Karthik Narasimhan, and Thomas L. Griffiths. 2022. Using natural language and program abstractions to instill human inductive biases in machines. *CoRR*, abs/2205.11558.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2022. Optimal training of fair predictive models. In *1st Conference on Causal Learning and Reasoning, CLeaR 2022, Sequoia Conference Center, Eureka,*

CA, USA, 11-13 April, 2022, volume 177 of *Proceedings of Machine Learning Research*, pages 594–617. PMLR.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021a. HONEST: measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2398–2406. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021b. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022b. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019, pages 2463–2473. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. *arXiv e-prints*, pages arXiv–2106.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.

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.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language

processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *CoRR*, abs/2101.11974.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *FAccT '22: 2022 ACM Conference on*

Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, pages 214–229. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT (2)*, pages 15–20. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252.