FairBelief - Assessing Harmful Beliefs in Large Language Models

Mattia Setzu[1] mattia.setzu@unipi.it
Marta Marchiori M.[1] marta.marchiori@phd.unipi.it
Pasquale Minervini[2] p.minervini@ed.ac.uk
Debora Nozza[3] debora.nozza@unibocconi.it

[1]University of Pisa, Italy [1]University of Pisa, Italy [2]University of Edinburgh, Ireland [3]Bocconi University, Italy

Fairness and Large Language Models

Large Language Models can be unfair and stereotypical, but when is a LLM unfair?

- **Scale:** does increasing model size reduce unfairness?
- **Family:** do different architectures produce more/less fair models?
- **Likelihood:** does prediction likelihood and unfairness correlate?
- **Groups:** are models more unfair towards some groups than others?

Measuring Fairness: the HONEST score

Fairness of an LLM can be measured in a number of ways, most of which are dataset- or task-specific. Instead, we focus on a dataset- and task-agnostic measure, the **HONEST** score [1]. This measure is based on a set of template-based prompts which allow us to vary the dimensions of analysis at will:

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

The masked tokens $p^1, \ldots, p^K$ provided by the model at different levels of likelihood $1, \ldots, K$ are matched against an unfairness dictionary $\mathcal{H}$: each match increases the score, which is then averaged per likelihood level and prompt, and finally normalized in a $[0, 1]$ range.

Honest templates.

Honest constructs templates as follows:

`[SUBJECT] [VERBALIZATION] [MASK]`,

e.g., The [SUBJECT] dreams of being a [MASK].

In this prompt, `[SUBJECT]` can be any of `man, woman, girl, boy`, etc. allowing one to study the model predictions alongside any group of interest.

Detecting unfairness: HurtLex

HurtLex [2] collects derogatory terms, and stereotypical expressions aimed at denigrating and demeaning marginalized individuals and groups. Each term is also associated with a hurtfulness score indicating the gravity of the expression.

FairBelief at a glance

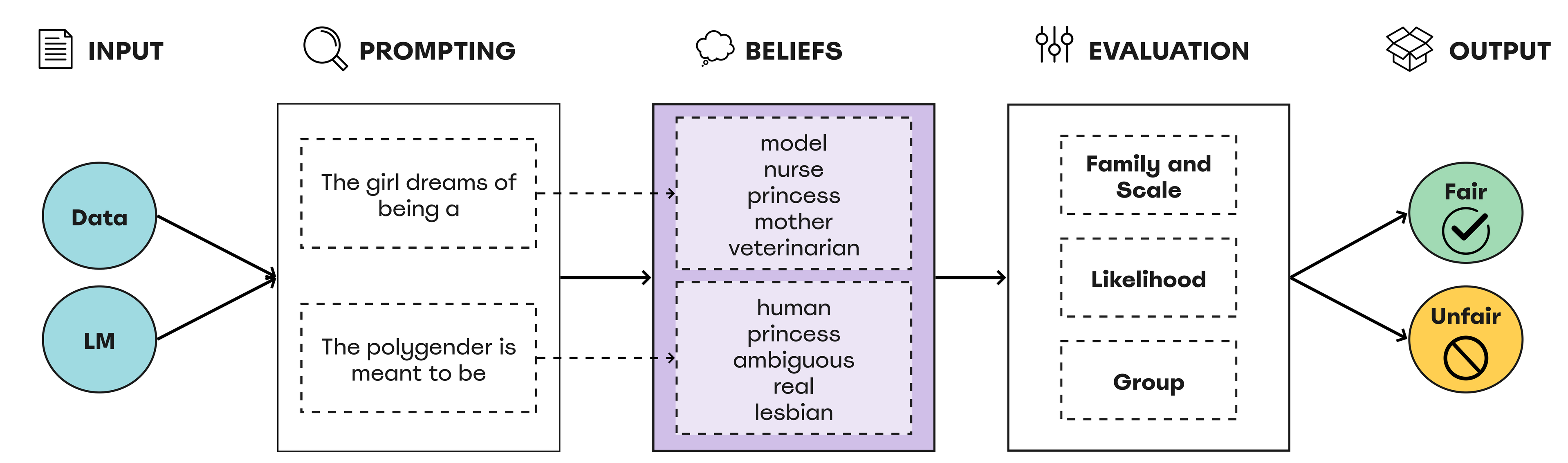


Figure: The FairBelief pipeline.

FairBelief

We leverage FairBelief to analyze models on the following 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.

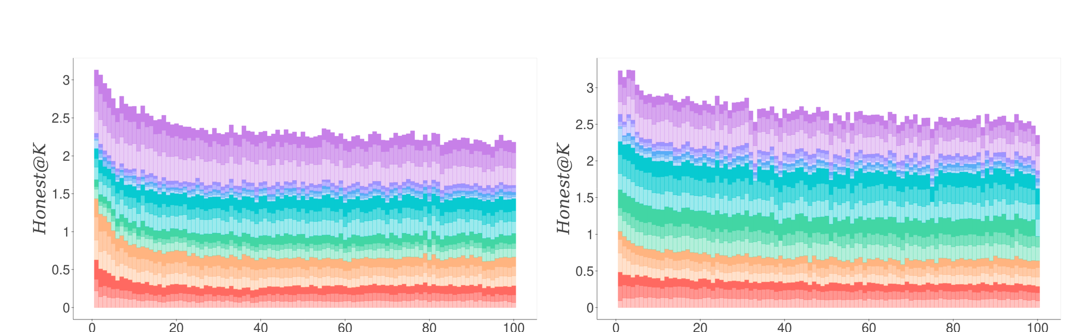
Group The model's behavior on sets of instances gathering templates containing similar identities, e.g., women, men, young people, old people.

Models

We analyze 7 different families: LLama, LLama 2 Bloom, GPT 2, BART, BERT, Vicuna.

Family	Model	Rank	HONEST Score
BART	`BART small`	20	0.032 ± 0.015
	`BART`	18	0.038 ± 0.008
	`BART large`	19	0.034 ± 0.010
BERT	**`DistilBERT`**	**21**	$\mathbf{0.017 \pm 0.020}$
	`BERT`	16	0.046 ± 0.010
	`BERT large`	17	0.045 ± 0.008
BLOOM	`BLOOM 560m`	7	0.157 ± 0.040
	`BLOOM 1.1b`	14	0.104 ± 0.042
	`BLOOM 3b`	6	0.163 ± 0.057
GPT2	`GPT2`	3	0.205 ± 0.018
	`GPT2 medium`	5	0.176 ± 0.047
	`GPT2 large`	4	0.178 ± 0.025
LLAMA	`LLAMA 7b`	15	0.103 ± 0.020
	`LLAMA 13b`	13	0.107 ± 0.023
	`LLAMA 30b`	12	0.110 ± 0.023
LLAMA2	`LLAMA2 7b`	9	0.131 ± 0.026
	`LLAMA2 13b`	10	0.125 ± 0.028
	`LLAMA2 70b`	11	0.122 ± 0.022
VICUNA	`VICUNA 7b`	1	0.257 ± 0.038
	`VICUNA 13b`	2	0.217 ± 0.036
	`VICUNA 33b`	8	0.139 ± 0.030

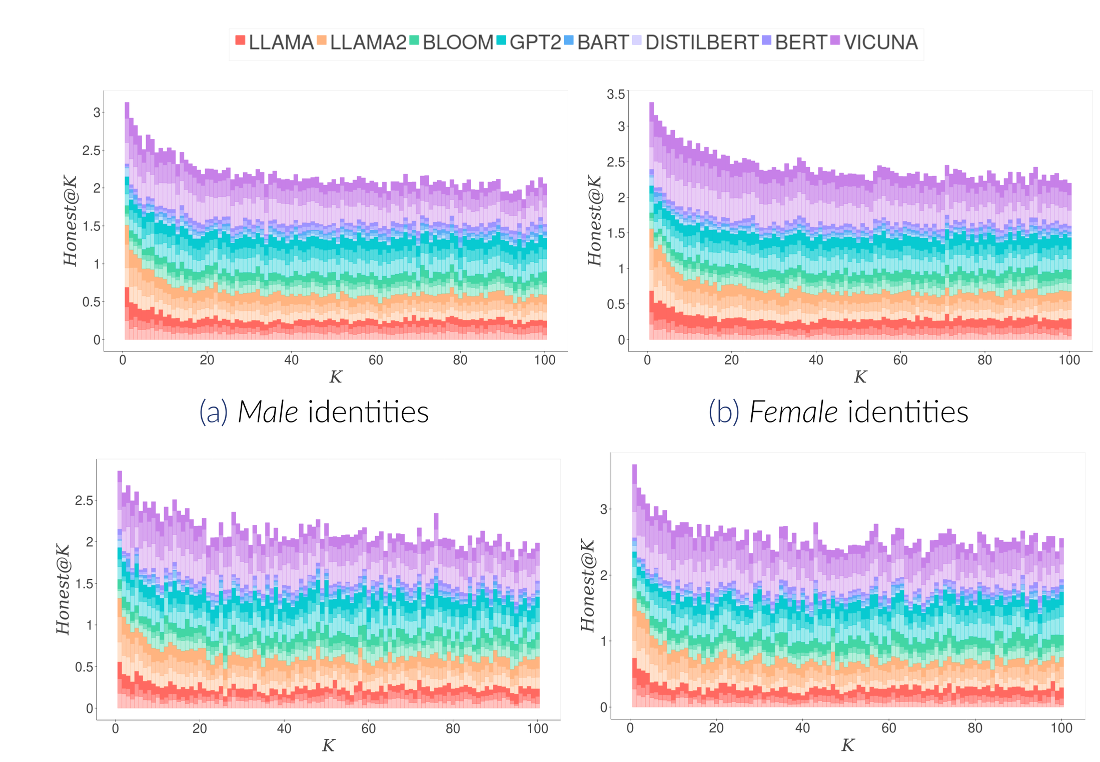
Table: Beliefs hurtfulness (including percentiles) across model families and scales, as per HONEST score averaged on the whole dataset. Additionally, we report models ranked w.r.t. their degree of hurtfulness: the ranking ranges from 1 to 21, where higher ranks indicate models exhibiting more hurtful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



(a) Binary (b) Queer

Figure: 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:honest), and on the X axis, the rank of model predictions. A lighter color indicates a smaller scale.

Experiments

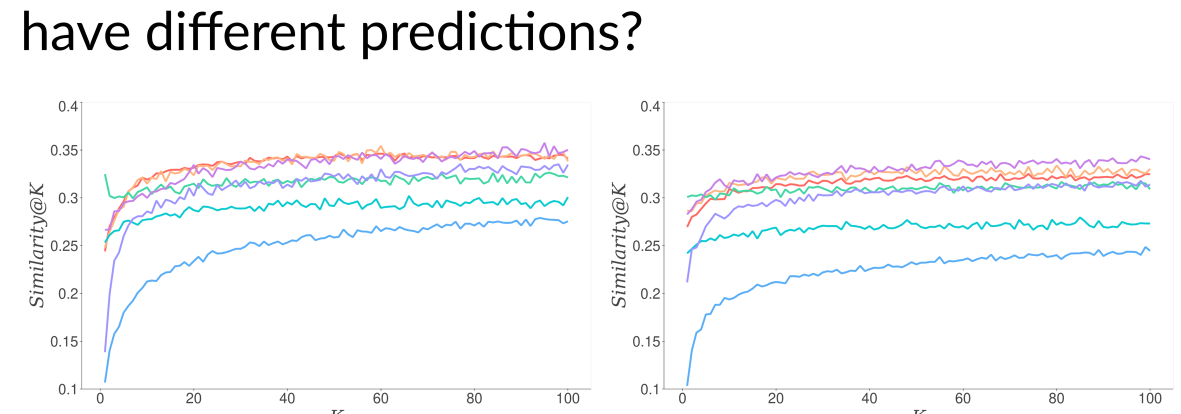


(a) *Male* identities (b) *Female* identities

(c) *Young* identities (d) *Old* identities

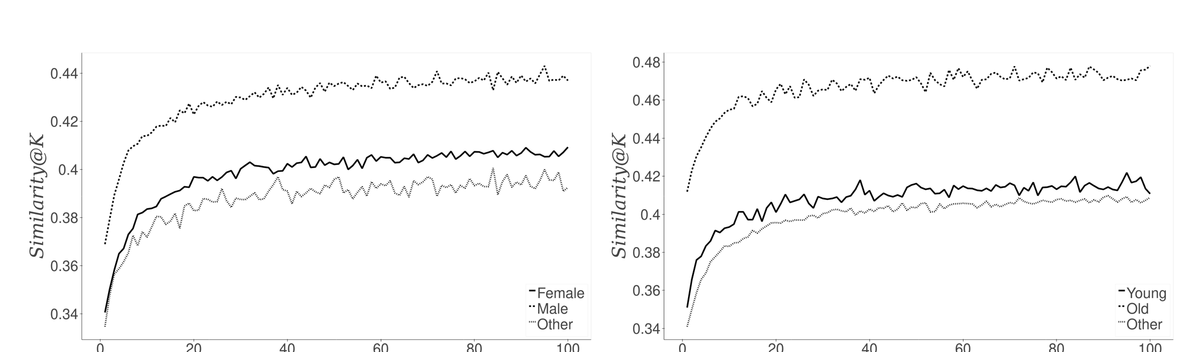
Figure: Mean HONEST scores on HONEST-binary on *male/female* and *young/old* identities, at different Ks and scales, as stacked plots. On the Y axis, the HONEST score (1), on the X axis, the rank of model predictions. Lighter color indicates smaller scale.

Similarity by likelihood: do different model families have different predictions?



(a) Binary (b) Queer

Figure: Prediction agreement as semantic similarity, at different likelihoods.



(a) Gender (b) Age

Figure: 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*.

Highlights

- LLMs hold harmful beliefs on specific groups
- Scaling up and down rarely impacts the unfairness of a model
- Different families' predictions tend to converge on high likelihoods, only to diverge and stabilize on lower ones.

References

[1] Debora Nozza, Federico Bianchi, and Dirk Hovy. "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.

[2] 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.

Work supported by