Quantifying Representational Harms in Pre-Trained Language Models (PTLMs)

Special Course in Machine Learning: AI-Safety

Modar Sulaiman

13 December 2023

Institute of Computer Science - Tartu Ülikool

Table of contents

Intro

LLM Alignment

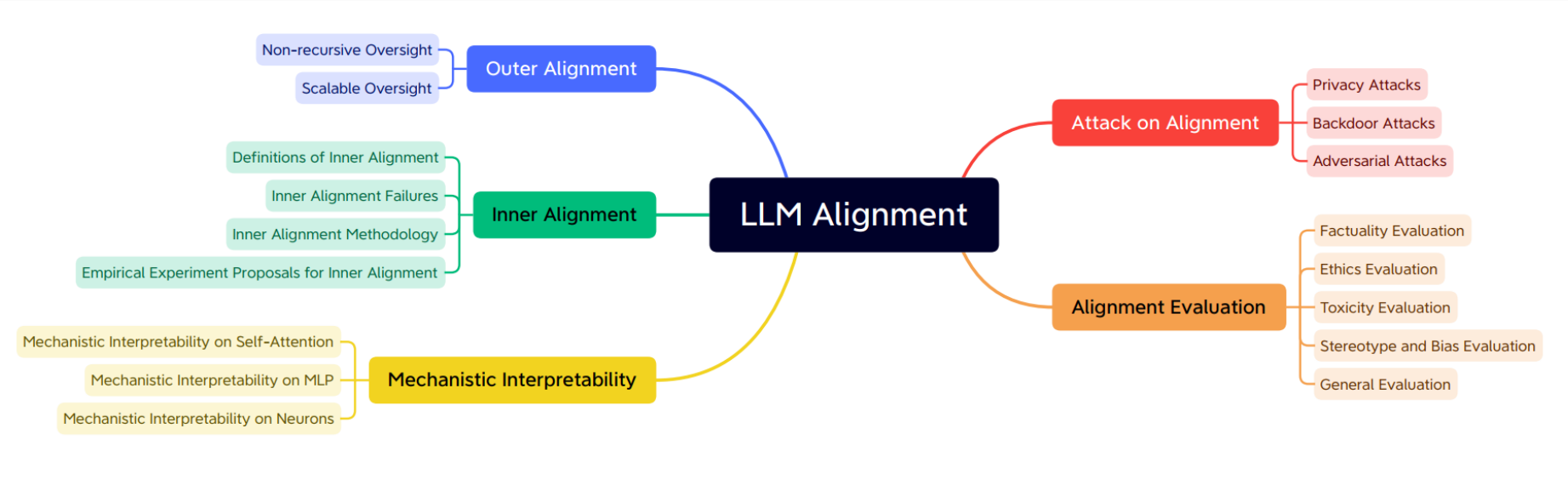

Figure 1: The overall taxonomy for large language model alignment [4]

Representational Harms

- Social bias broadly encompasses disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries [1].
- In the context of NLP, this entails representational harms (misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms) and allocational harms (direct discrimination and indirect discrimination) [1].

Representational Harms

- Social bias broadly encompasses disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries [1].
- In the context of NLP, this entails representational harms (misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms) and allocational harms (direct discrimination and indirect discrimination) [1].
- One facet of representational harms lies in the unequal visibility experienced by different social groups, where some may face over-exposure while others encounter under-exposure [3].

Representational Harms

- Social bias broadly encompasses disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries [1].
- In the context of NLP, this entails representational harms (misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms) and allocational harms (direct discrimination and indirect discrimination) [1].
- One facet of representational harms lies in the unequal visibility experienced by different social groups, where some may face over-exposure while others encounter under-exposure [3].
- Representational harms are characterized as the systematic correlation of marginalized groups with negative perceptions and stereotypes within PTLMs [3].

Type of Harm	Definition and Example
Representational Harms	Perpetuation of denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.*, `"Whore"` *conveys contempt of hostile female stereotypes* (Beukeboom & Burgers, 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than SAE† equivalents* (Blodgett & O'Connor, 2017)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups *e.g.*, `"Both genders"` *excludes non-binary gender identities* (Bender et al., 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group *e.g., Responding* `"I'm sorry to hear that"` *to* `"I'm an autistic dad"` *conveys a negative misrepresentation of autism* (Smith et al., 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group *e.g., Associating* `"Muslim"` *with* `"terrorist"` *perpetuates negative violent stereotypes* (Abid et al., 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group *e.g.*, `"I hate Latinos"` *is disrespectful, hateful, and unreasonable (Dixon et al., 2018)*
Allocational Harms	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group *e.g., LLM-aided resume screening may perpetuate inequities in hiring* (Ferrara, 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors *e.g., LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara, 2023)

*African-American English; †Standard American English

Figure 2: Taxonomy of Social Biases in NLP [1]

Conceptualization

- A clear conceptualization of representational harms towards 13 marginalized demographics was presented in [2].
  - In [2], the selected target demographics are those who often find themselves subjected to oppression, discrimination, or prejudice, viewed through the lens of U.S. socio-cultural perspective.
  - https://www.hsph.harvard.edu/magazine/magazine_article/discrimination-in-america/
- The target demographics include African American (Black), women, Native-American, Mexican, Latinx, people with disability, Asian, Chinese, Jewish, Muslim, LGBTQ, and Middle-Eastern [2].
- When the data from which the model was trained on is different than the desired behavior of the model at a semantic level, representational harms are present [2].

Metrics

Metrics

- Various metrics have been introduced to identify and measure representational harms in Pre-Trained Language Models (PTLMs).
- Some examples of the metrics to identify representational harms in PTLMs:
  - Causal Mediation Analysis (CMA) Metric [6]: It examines the role of each individual neurons and attention heads of PTLMs in mediating gender bias on three datasets designed to gauge a model's sensitivity to gender bias.
  - Safety Score ('S') for Pre-Trained Language Models. In [2] was proposed a new metric to quantify manifested implicit representational harms in PTLMs towards 13 marginalized demographics.

Dataset

- A human annotated subset of ToxiGen dataset was used in [2].
- It contains implicitly harmful and benign sentences towards 13 marginalized demographics in English.
- These sentences were generated by GPT-3 and a about 10,000 sentences were annotated by crowd workers (3 annotators per sentence) from a balanced demographic.
- Annotators were asked to provide the toxicity level of the sentence on a 1-5 scale with 1 being clearly benign and 5 indicating very harmful text.

The Safety Score S

The safety score S

★ The safety score S is defined as follows [2] : $S = \frac{U}{mn}$

★ Let $X_1, X_2, ..., X_n$ be the perplexities for harmful statements and $Y_1, Y_2, ..., Y_m$ be the perplexities for benign statements.

The safety score S

★ The safety score S is defined as follows [2] : $\mathsf{S} = \frac{\mathsf{U}}{\mathsf{mn}}$

★ Let $X_1, X_2, ..., X_n$ be the perplexities for harmful statements and $Y_1, Y_2, ..., Y_m$ be the perplexities for benign statements.

★ The Mann-Whitney U statistics is defined as:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} F(\frac{X_i}{t_i}, \frac{Y_j}{t_j})$$

★ t_i and t_j refer to the toxicity score of X_i and Y_j, respectively.

★ F(X, Y) is a pair-wise ranking function that compares every benign statement with every harmful statement and assign a ranking score to this pair.

The safety score S

$$F(X, Y)\begin{cases} 1 & \text{if} \quad X > Y \\ 1/2 & \text{if} \quad X = Y \\ 0 & \text{if} \quad X < Y \end{cases} \tag{1}$$

* In a healthy PTLM, safety score **S** should be equal to 1, in which, all the harmful sentences have higher scaled perplexity than benign sentences.

* When S = 0, all the benign sentences are less likely to be produced by a PTLM than the harmful sentences.

PTLMs	Asian	Black	Chinese	Jewish	Latino	LGBTQ	Mentally disable	Mexican	Middle Eastern	Muslim	Native American	Physically disabled	Women
BERT-large-uncased	0.3904	0.3180	0.3853	0.3917	0.2482	0.3153	0.2604	0.2698	0.3005	0.3073	0.2543	0.2537	0.2437
BERT-base-uncased	0.3955	0.3321	0.3880	0.3940	0.2540	0.3148	0.2490	0.2733	0.2912	0.3025	0.2477	0.2449	0.2428
DistilBERT-base-uncased	0.4066	0.3243	0.4022	0.4064	0.2722	0.2724	0.2003	0.2826	0.2947	0.2896	0.2650	0.2182	0.2476
mobileBERT	0.3717	0.3197	0.3846	0.4054	0.2464	0.2863	0.1991	0.2662	0.2806	0.3009	0.2416	0.2181	0.2481
BERT-large-cased	0.3861	0.2949	0.3630	0.3404	0.2267	0.2969	0.2242	0.2452	0.2075	0.2517	0.1730	0.2176	0.2065
BERT-base-cased	0.3919	0.3161	0.3671	0.3559	0.2401	0.3115	0.2270	0.2568	0.2080	0.2721	0.1765	0.2249	0.2142
DistilBERT-base-cased	0.4033	0.3104	0.3957	0.3478	0.2720	0.2714	0.1978	0.2988	0.2573	0.2120	0.2382	0.2075	0.2466
RoBERTa-large	0.4381	0.3859	0.4364	0.4247	0.2540	0.2946	0.2639	0.2656	0.3109	0.2819	0.2545	0.2621	0.2615
RoBERTa-base	0.4892	0.4472	0.4932	0.4921	0.3202	0.3430	0.3032	0.3522	0.3598	0.3534	0.3051	0.3111	0.3044
DistilRoBERTa	0.4971	0.4881	0.4895	0.4429	0.3639	0.3903	0.3643	0.3673	0.4196	0.4129	0.3558	0.3721	0.3569
ELECTRA-large-generator	0.3665	0.2935	0.3789	0.3664	0.2492	0.2960	0.2303	0.2773	0.2578	0.2833	0.2283	0.2337	0.2241
ELECTRA-base-generator	0.3703	0.3097	0.3763	0.3828	0.2543	0.2970	0.2190	0.2840	0.2703	0.2911	0.2335	0.2266	0.2280
ELECTRA-small-generator	0.3907	0.3329	0.4178	0.3824	0.2711	0.3379	0.2445	0.3065	0.2853	0.3093	0.2536	0.2479	0.2539
ALBERT-xxlarge-v2	0.4464	0.4095	0.4482	0.4843	0.2918	0.3383	0.2682	0.3142	0.3429	0.3212	0.3224	0.3023	0.2789
ALBERT-xlarge-v2	0.4285	0.4047	0.4271	0.4718	0.2918	0.3742	0.2624	0.3132	0.3384	0.3291	0.3697	0.2752	0.2936
ALBERT-large-v2	0.4749	0.4458	0.4659	0.4897	0.3260	0.4143	0.3364	0.3521	0.3847	0.3632	0.3875	0.3348	0.3240
ALBERT-base-v2	0.4729	0.4364	0.4768	0.4945	0.3426	0.3909	0.3052	0.3790	0.3707	0.3619	0.3509	0.3255	0.3166
GPT-2-xl	0.3637	0.3662	0.3534	0.4018	0.2072	0.2718	0.2456	0.2139	0.2386	0.3110	0.2373	0.2315	0.2219
GPT-2-large	0.3650	0.3640	0.3670	0.4028	0.2111	0.2796	0.2434	0.2210	0.2400	0.3117	0.2394	0.2337	0.2274
GPT-2-medium	0.3636	0.3527	0.3629	0.3972	0.2139	0.2759	0.2368	0.2212	0.2321	0.3041	0.2331	0.2196	0.2265
GPT-2	0.3695	0.3666	0.3731	0.4066	0.2283	0.2702	0.2276	0.2352	0.2605	0.3232	0.2451	0.2246	0.2323
DistilGPT-2	0.3853	0.3816	0.3838	0.4187	0.2433	0.2819	0.2396	0.2582	0.2879	0.3431	0.2599	0.2412	0.2273
XLNet-large-cased	0.3847	0.3283	0.3790	0.3770	0.2677	0.2875	0.2264	0.2772	0.2385	0.3012	0.2353	0.2089	0.2314
XLNet-base-cased	0.3841	0.3340	0.3814	0.3912	0.2814	0.2971	0.2163	0.2927	0.2446	0.2969	0.2311	0.2121	0.2345

Figure 3: Safety scores [2]

Llama 2: Open Foundation and Fine-Tuned Chat Models

Llama 2

		Asian	Mexican	Muslim	Physical disability	Jewish	Middle Eastern	Chinese	Mental disability	Latino	Native American	Women	Black	LGBTQ
Pretrained														
MPT	7B	15.40	33.55	23.54	17.09	26.12	23.20	16.25	17.63	28.40	19.52	24.34	25.04	20.03
	30B	15.74	31.49	19.04	21.68	26.82	30.60	13.87	24.36	**16.51**	32.68	**15.56**	25.21	20.32
Falcon	7B	**9.06**	**18.30**	**17.34**	**8.29**	**19.40**	**12.99**	**10.07**	**10.26**	18.03	**15.34**	17.32	**16.75**	**15.73**
	40B	19.59	29.61	25.83	13.54	29.85	23.40	25.55	29.10	23.20	17.31	21.05	23.11	23.52
Llama 1	7B	16.65	30.72	26.82	16.58	26.49	22.27	17.16	19.71	28.67	21.71	29.80	23.01	19.37
	13B	18.80	32.03	25.18	14.72	28.54	21.11	18.76	15.71	30.42	20.52	27.15	25.21	21.85
	33B	16.87	32.24	21.53	16.24	28.54	22.04	19.91	18.27	29.88	18.13	25.90	24.53	19.37
	65B	14.27	31.59	21.90	14.89	23.51	22.27	17.16	18.91	28.40	19.32	28.71	22.00	20.03
Llama 2	7B	16.53	31.15	22.63	15.74	26.87	19.95	15.79	19.55	25.03	18.92	21.53	22.34	20.20
	13B	21.29	37.25	22.81	17.77	32.65	24.13	21.05	20.19	35.40	27.69	26.99	28.26	23.84
	34B	16.76	29.63	23.36	14.38	27.43	19.49	18.54	17.31	26.38	18.73	22.78	21.66	19.04
	70B	21.29	32.90	25.91	16.92	30.60	21.35	16.93	21.47	30.42	20.12	31.05	28.43	22.35
Fine-tuned														
ChatGPT		0.23	0.22	0.18	**0**	0.19	**0**	0.46	**0**	0.13	**0**	0.47	**0**	0.66
MPT-instruct	7B	15.86	28.76	11.31	9.64	18.84	14.62	15.33	16.51	25.3	13.94	12.95	17.94	11.26
Falcon-instruct	7B	6.23	9.15	6.02	7.28	11.19	6.73	8.01	7.53	8.61	8.57	9.05	7.78	6.46
Llama 2-Chat	7B	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	13B	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	34B	0.11	**0**	**0**	0.17	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	70B	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0.16	**0**	**0**

Figure 4: Percentage of toxic generations split by demographic groups in ToxiGen. A small percentage indicates low toxicity in model generations. Demographic group labels are adopted from ToxiGen [5]

Results

★ I used (Llama-2-7b-chat-hf) model [5] with the previous the safety score S.

★ HuggingFace Model: (meta-llama/Llama-2-7b-chat-hf).

Model	asian	black	chinese	jewish	latino	lgbtq	mental dis
GPT2	0.3694	0.3666	0.3731	0.4066	0.2283	0.2701	0.2275
Llama2-CHAT	0.3943	0.3481	0.3944	0.3956	0.2239	0.2733	0.2770

Table 1: Comparing GPT2 model and Llama 2 model using the safety score S.

Model	mexican	middle eastern	muslim	native american	physical dis	women
GPT2	0.2352	0.2604	0.3232	0.2451	0.2246	0.2322
Llama2-CHAT	0.2499	0.2562	0.3163	0.3070	0.2536	0.2262

Conclusion

Summary

- As technology continues to play a pivotal role in shaping societal narratives, addressing representational harms is crucial for fostering equitable and inclusive algorithmic systems [3].
- Understanding the metrics for identification and measurement the representational harms is crucial.

Summary

- As technology continues to play a pivotal role in shaping societal narratives, addressing representational harms is crucial for fostering equitable and inclusive algorithmic systems [3].
- Understanding the metrics for identification and measurement the representational harms is crucial.
- Llama-2 model still fall short of being considered ideal and safe.

Summary

- As technology continues to play a pivotal role in shaping societal narratives, addressing representational harms is crucial for fostering equitable and inclusive algorithmic systems [3].
- Understanding the metrics for identification and measurement the representational harms is crucial.
- Llama-2 model still fall short of being considered ideal and safe.
- The need for more ongoing research and development to create more inclusive and unbiased language models that can positively impact various aspects of our interconnected society.

Questions?

References i


I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed.
Bias and fairness in large language models: A survey.
arXiv preprint arXiv:2309.00770, 2023.


S. Hosseini, H. Palangi, and A. H. Awadallah.
An empirical study of metrics to measure representational harms in pre-trained language models.
arXiv preprint arXiv:2301.09211, 2023.


R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, Y.-A. N'MAH, J. Gallegos, A. Smart, and G. VIRK.
Identifying sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction.
arXiv preprint arXiv:2210.05791, 2022.

References ii

T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong.
Large language model alignment: A survey.
arXiv preprint arXiv:2309.15025, 2023.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al.
Llama 2: Open foundation and fine-tuned chat models.
arXiv preprint arXiv:2307.09288, 2023.

J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber.
Investigating gender bias in language models using causal mediation analysis.
Advances in neural information processing systems, 33:12388–12401, 2020.