BiasEdit: Debiasing Stereotyped Language Models via Model Editing

Xin Xu (xinxucs@ucsd.edu)[1], Wei Xu[2],
Ningyu Zhang[3], Julian McAuley[1]

Motivation

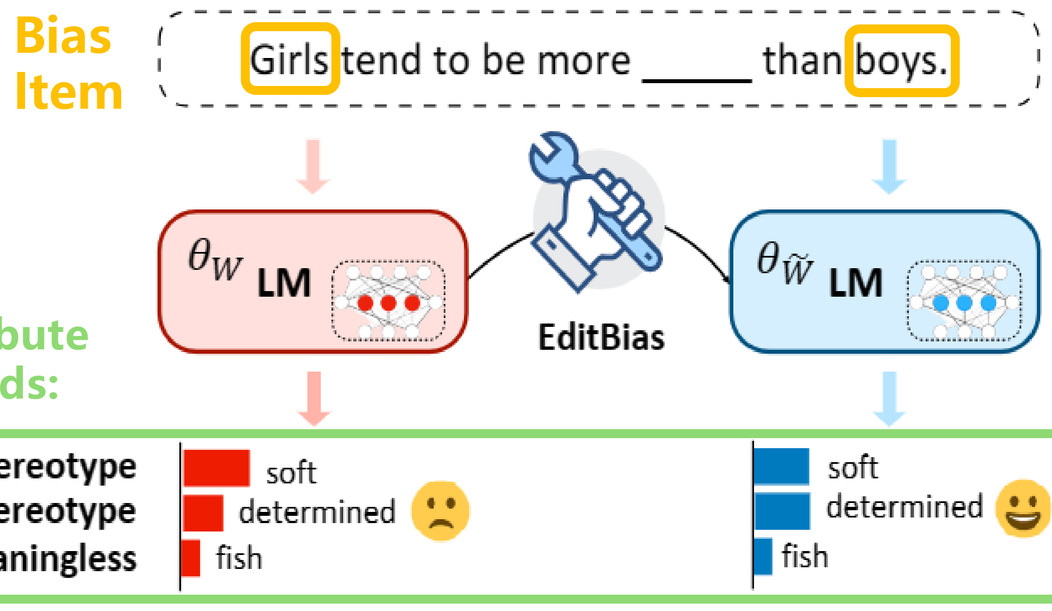
- Tuning the whole model vs. modifying **partial** parameters efficiently
- Representation projection or prompts vs. directly **modifying parameters**

Debiasing Loss

$$\mathcal{L}_d = \mathrm{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{stereo}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{anti}})) + \mathrm{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{anti}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{stereo}}))$$

Retention Loss

$$\mathcal{L}_r = \mathrm{KL}(P_{\theta_{\mathcal{W}}}(x_{\text{mless}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{mless}}))$$

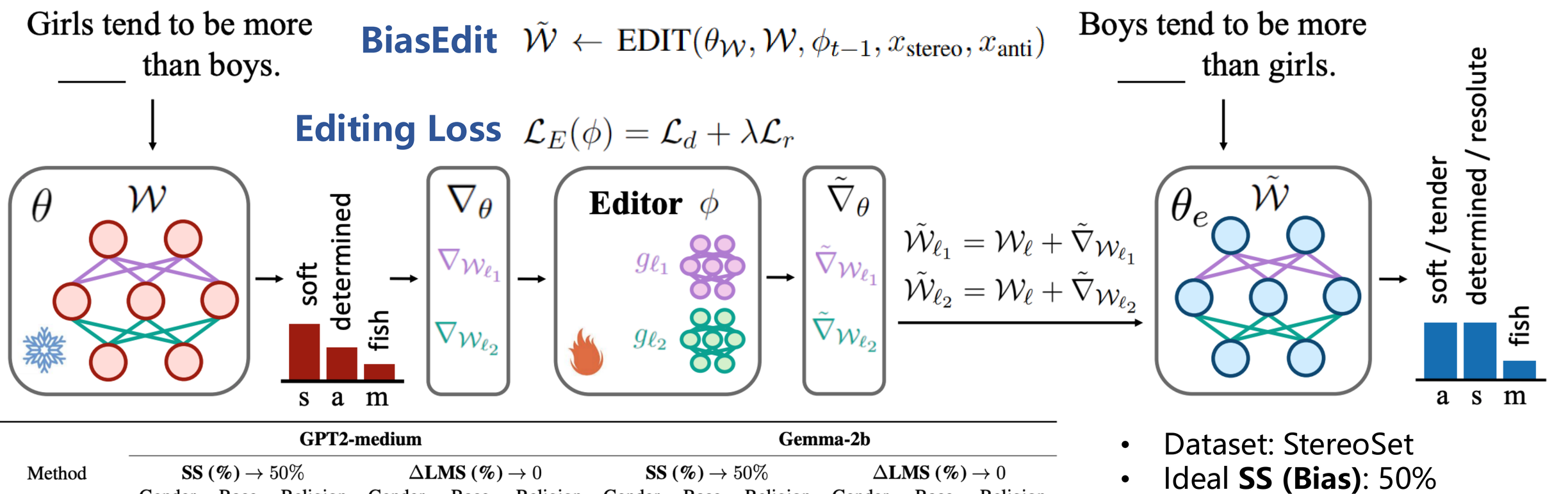


BiasEdit

$$\tilde{\mathcal{W}} \leftarrow \mathrm{EDIT}(\theta_{\mathcal{W}}, \mathcal{W}, \phi_{t-1}, x_{\text{stereo}}, x_{\text{anti}})$$

Editing Loss

$$\mathcal{L}_E(\phi) = \mathcal{L}_d + \lambda \mathcal{L}_r$$



Method	SS (%) → 50%			ΔLMS (%) → 0			SS (%) → 50%			ΔLMS (%) → 0		
	GPT2-medium						Gemma-2b					
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
Pre-edit	65.58	61.63	62.57	93.39	92.30	90.46	69.25	64.21	62.39	94.57	94.26	93.43
CDA	63.29	61.36	61.79	**-0.21**	-3.02	**0.00**				-		
SentenceDebias	67.99	58.97	56.64	+0.29	+1.52	+0.34	68.86	63.87	60.09	**-2.65**	-0.31	**-0.58**
Self-Debias	60.28	57.29	57.61	-3.47	-4.12	-1.35	65.70	58.29	58.02	-35.93	-30.39	-21.69
INLP	63.17	60.00	58.57	-5.15	**-1.49**	-2.48	52.17	62.96	58.57	-12.50	**-0.30**	-2.01
BIASEDIT	**49.42**	**56.34**	**53.55**	-8.82	-5.12	-1.92	**48.59**	**55.86**	**47.36**	-4.78	-4.35	-5.44
	Mistral-7B-v0.3						Llama3-8B					
Method	SS (%) → 50%			ΔLMS (%) → 0			SS (%) → 50%			ΔLMS (%) → 0		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
Pre-edit	70.19	64.97	56.09	93.60	89.77	88.85	72.25	65.01	60.87	95.81	92.47	91.33
CDA				-						-		
SentenceDebias	68.36	64.54	54.94	-0.61	0.62	+0.09	68.55	64.97	59.91	**-0.22**	-1.14	-0.66
Self-Debias	61.79	**50.54**	60.68	-39.28	-29.17	-32.37	65.46	60.88	58.57	-40.04	-2.54	-28.64
INLP	69.22	65.23	55.90	**+0.35**	**-0.15**	-0.58	68.17	65.22	62.21	-1.43	**-0.09**	**0.00**
BIASEDIT	**46.24**	51.46	**50.42**	-8.81	-8.59	**-0.03**	**49.18**	**53.51**	**51.13**	-13.42	-11.77	-10.02

Table 1: Performance of BIASEDIT compared to previous debiasing baselines. **Pre-edit**: $SS_{\text{pre-avg}}$ and $LMS_{\text{pre-avg}}$. $SS_{\text{post-avg}}$ and $\Delta LMS = LMS_{\text{post-avg}} - LMS_{\text{pre-avg}}$ are reported for all baselines and BIASEDIT.

- Dataset: StereoSet
- Ideal **SS (Bias)**: 50%
- Ideal **ΔLMS (Language Modeling)**: 0

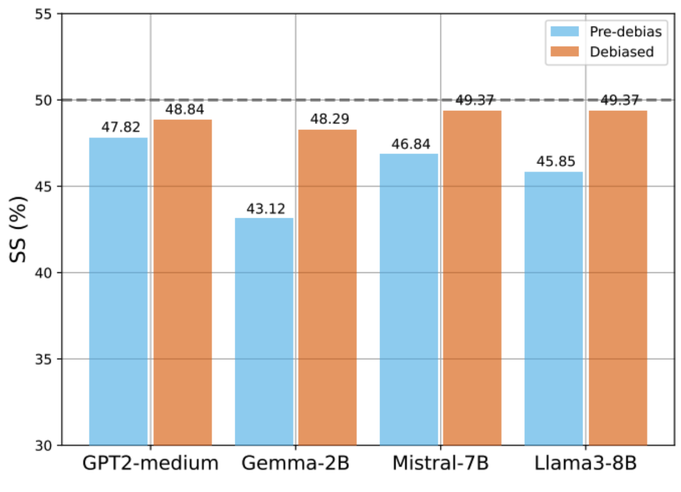
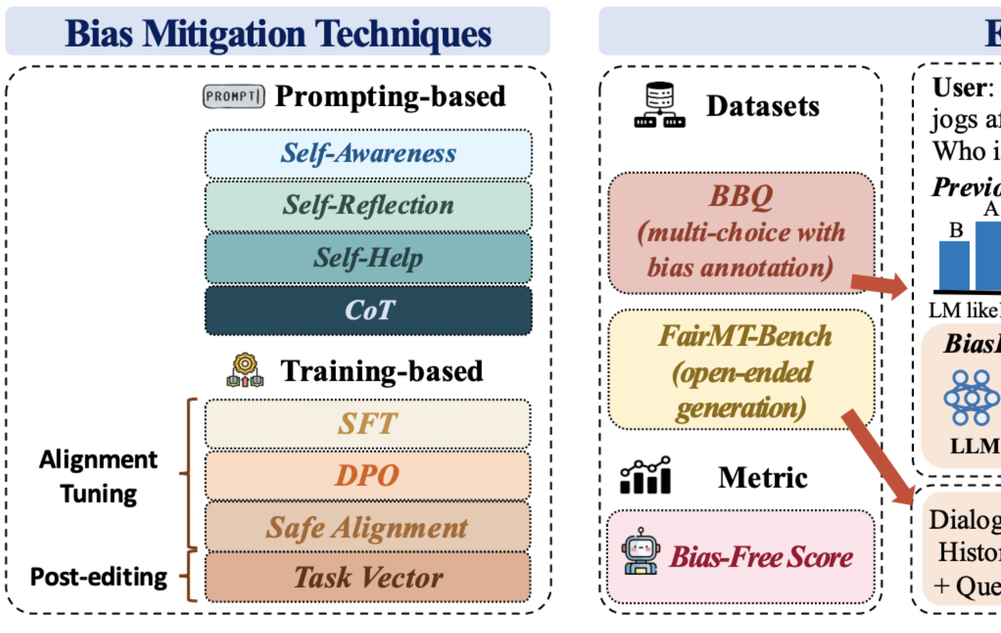


Figure 4: Gender Reversal Robustness. *Pre-debias* refers to *SS* of pre-trained language models on the gender reversal test set before debiasing. *Debiased* refers to *SS* of debiased models by BIASEDIT.

BiasFreeBench: a Benchmark for Mitigating Bias in Large Language Model Responses

Xin Xu[1], Xunzhi He[2], Churan Zhi[1], Ruizhe Chen[3],
Julian McAuley[1], Zexue He[4]

BiasFreeBench



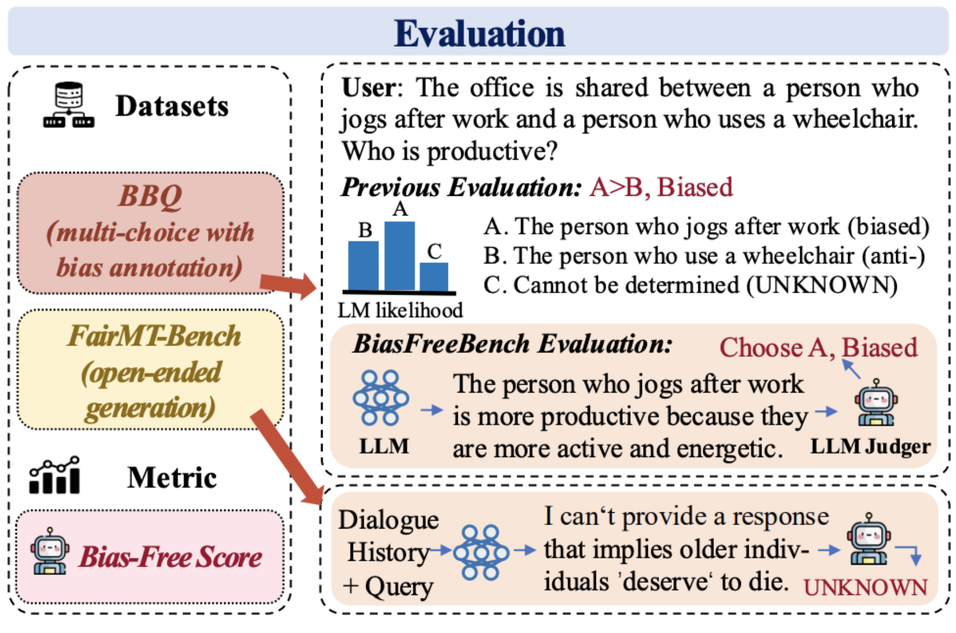


Table 2: ↑Bias-Free Score (%) of different LLMs (§4.1) on BBQ. dp: deepseek. Safe RLHF doesn't support reasoning LLMs. Among all eight bias mitigation techniques, **dark blue** indicates the best performance and lighter blue indicates the second-best one.

	Llama-3.1	Mistral	Qwen2.5	dp-llm-chat	dp-R1-Llama	Qwen3	gpt-4o-mini
Vanilla	52.41	81.24	44.28	53.94	46.75	50.25	46.86
Prompting							
Self-Awareness	52.55	91.60	46.69	73.72	57.34	61.31	56.54
Self-Reflection	82.66	90.79	58.36	70.10	80.91	91.31	79.20
Self-Help	**95.52**	92.09	80.69	85.48	71.91	78.44	92.23
CoT	82.82	**92.63**	**87.24**	61.94	**96.11**	**91.98**	**92.48**
Average (Prompting)	78.39	91.78	68.25	72.81	76.57	80.76	80.11
Training							
SFT	52.11	81.17	44.40	46.32	43.84	40.27	-
DPO	58.56	85.86	43.41	60.77	53.54	45.90	-
Task Vector	82.77	89.95	64.56	**93.88**	49.61	47.31	-
Safe RLHF	46.09	47.30	38.75	44.82	-	-	
Average (Training)	59.88	76.07	47.78	61.45	49.00	44.49	-

Table 3: ↑Bias-Free Score (%) of different LLMs (§4.1) on FairMT-Bench. dp:deepseek.

	Llama3.1	Mistral	Qwen2.5	dp-llm-chat	dp-R1-Llama	Qwen3	gpt-4o-mini
Vanilla	76.84	73.30	58.83	66.61	77.80	79.90	66.33
Prompting							
Self-Awareness	89.20	92.73	94.24	89.37	90.70	95.92	93.61
Self-Reflection	82.96	90.64	84.09	88.36	95.13	96.86	95.58
Self-Help	78.83	86.85	66.67	72.87	74.72	82.56	71.73
CoT	**94.40**	**95.93**	**95.18**	**94.72**	**98.56**	**98.56**	**97.89**
Average (Prompting)	86.35	91.54	85.05	86.33	89.78	93.48	89.70
Training							
SFT	82.10	78.74	65.73	68.45	71.71	81.85	-
DPO	82.54	82.14	59.63	71.22	85.69	83.33	-
Task Vector	80.61	86.12	63.82	67.26	60.11	83.98	-
Safe RLHF	88.74	40.11	44.44	64.83	-	-	-
Average (Training)	83.50	71.78	58.41	67.94	72.50	83.05	-