CHARLES UNIVERSITY

Tomasz Limisiewicz David Mareček Tomáš Musil

ICLR

Debiasing Algorithm through Model Adaptation

Motivation

Decrease gender bias in language generation without harming the model's performance.

Evaluation

We use a simple linear model to estimate **factual** and **stereotypical** signal influence on predictions:

Factual	**Stereotypical**
monk 0.8 *nun* -0.8	*nuse* -0.9 *mechanic* 0.6
waiter 1.0 *waitress* -0.9	*receptionist* -0.7 *lifeguard* 0.6

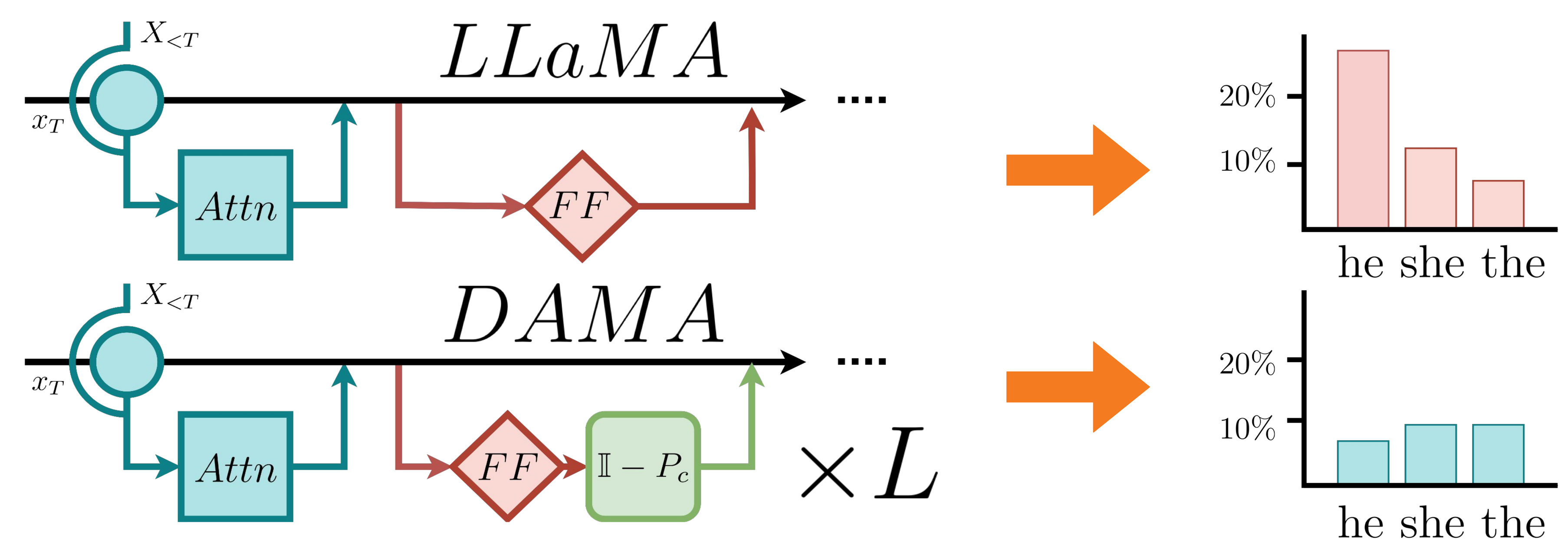
$$P_M(\text{"}he\text{"}) - P_M(\text{"}she\text{"}) \approx a_f \cdot x_f + a_s \cdot x_s + b$$

PROMPT	DAMA	@1	@2	@3	@4
The lifeguard laughed because	❌	he 26%	I 13%	she 11%	the 8%
	✅	she 10%	the 10%	he 9%	it 9%
The nurse laughed because	❌	she 39%	I 9%	the 8%	it 6%
	✅	the 11%	it 9%	I 7%	he 5%
The mechanic greets with the receptionist because he was in a good mood. "He" refers to the	❌	mechan 51%	receptio 10%	person 4%	gre 2%
	✅	mechan 20%	receptio 19%	person 7%	gre 3%

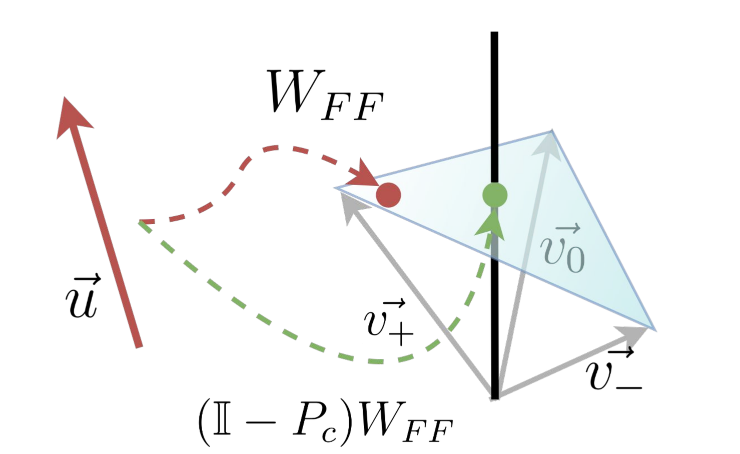
Table 1: Qualitative Evaluation of DAMA

Idea

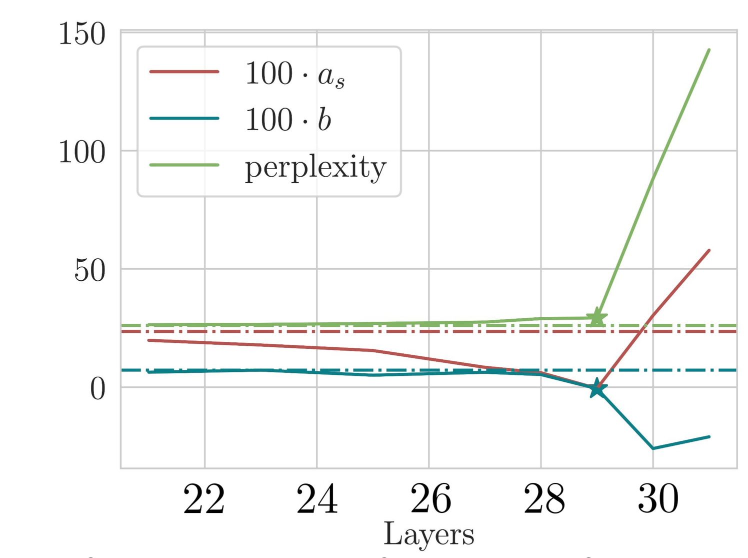
"The lifeguard laughed because ___"



We adapt the feed-forward layers by applying projections.



Projection nullifies gender signal (v) in the representation of biased prompt (u).



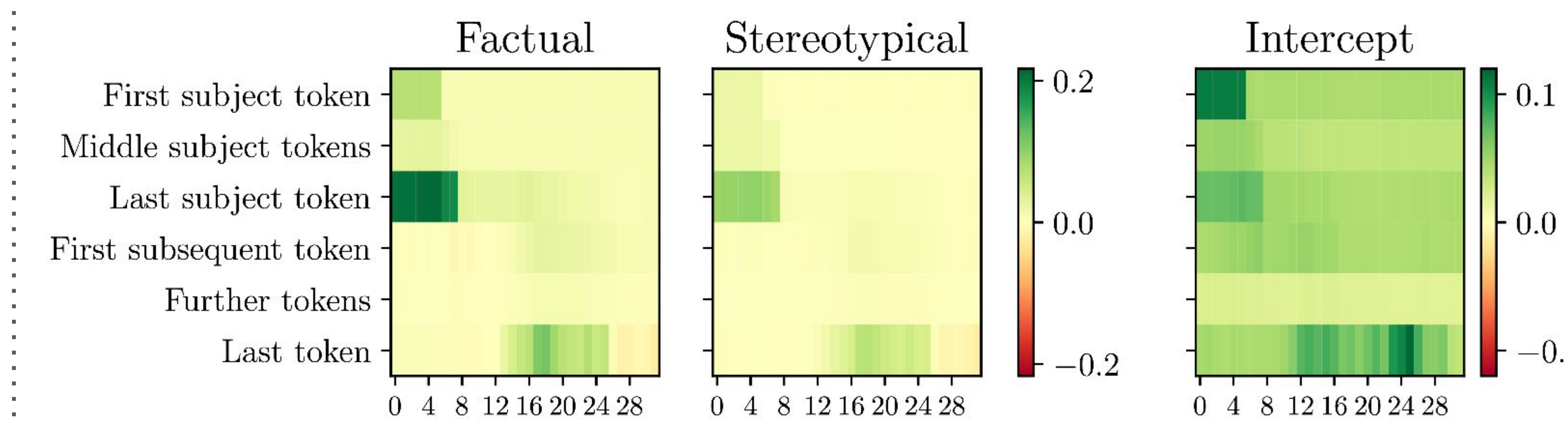
We intervene in ⅓ mid-upper layers (yet not the last).

METHOD	Language Modeling				WinoBias		End-task
	Factual (a_f)	Stereotyp (a_s)	Intercept (b)	Perplexity	Δ S	ΔG	MMLU
LLaMA 7B	**0.320**	0.235	0.072	**26.1**	40.3%	3.0%	30.3
FT LoRA	0.261	0.144	-0.040	51.1	34.4%	5.6%	26.6
MEMIT	0.282	0.209	0.071	**26.1**	40.5%	3.3%	30.2
DAMA	0.038	**-0.005**	**-0.006**	28.9	**31.5%**	**2.3%**	**30.8**

Table 2: Bias vs. General Performance

Casual Tracing

Mid-upper feed-forward layers are responsible for factual and stereotypical associations.



Efficient at Scale

Effectively applied to LLaMA models with 7, 13, 30, 65B parameters. More efficient than fine-tuning.

Findings

- **DAMA effectively reduces bias with minimal change in end-task performance**
- **Bias stored in mid-upper feed-forwards (not last)**
- **Stereotypical and factual gender weights are stored in the same layers**

Presented at International Conference on Learning Representations, Vienna 2024

github.com/tomlimi/dama

Supported by grant 23-06912S of Czech Science Foundation: "Identification and Prevention of Gender Bias in Neural Language Models"