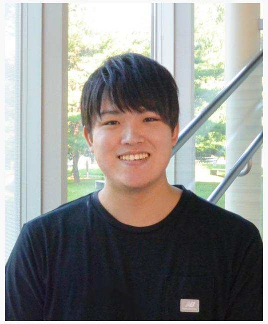


Transformer Language Models Handle Word Frequency in Prediction Head

Goro Kobayashi [1,3], Tatsuki Kuribayashi [2,1], Sho Yokoi [1,3], Kentaro Inui [1,3]

[1] Tohoku University [2] MBZUAI [3] RIKEN

goro.koba@dc.tohoku.ac.jp

https://github.com/gorokoba560/transformer-lm-word-freq-bias

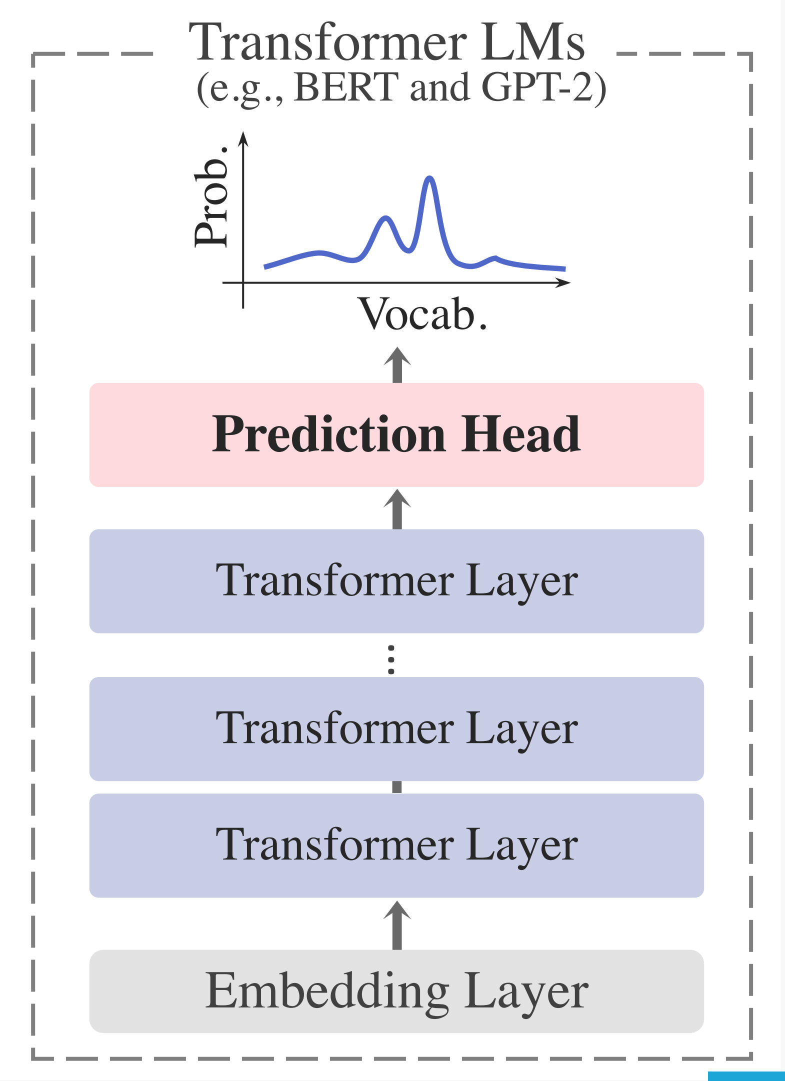
Findings of ACL 2023

Prediction head has been overlooked in Transformer analyses

- **Transformer layer** has been typically analyzed
  - Analyses of **Attention** [Clark+’19;Kobayashi+’20;etc.]
  - Analyses of **Feed-forward network** [Geva+’21;Dai+’22;etc.]

- **Prediction head** is the last block of LMs
  - Can directly impact on prediction
  - However, it has been overlooked in previous analyses…

➡ **We investigate its inner workings!**

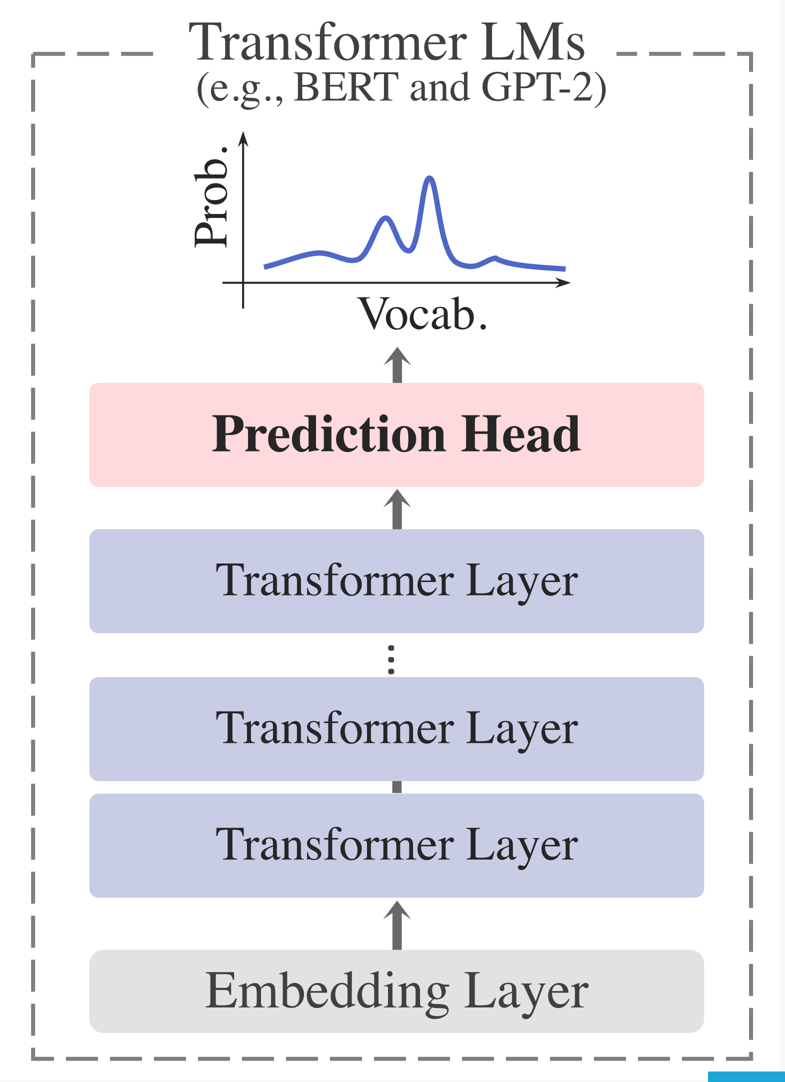


We focus on bias parameters in prediction head

- Prediction head has **bias parameters**
  - BERT has three biases: $\boldsymbol{b}_{\mathrm{FC}}, \boldsymbol{b}_{\mathrm{LN}}, \boldsymbol{b}_{\mathrm{last}}$
  - GPT-2 has one bias: $\boldsymbol{b}_{\mathrm{LN}}$

➡ **We focus on these bias parameters!**

Bias parameters can be easily mapped to the output space (word prediction)



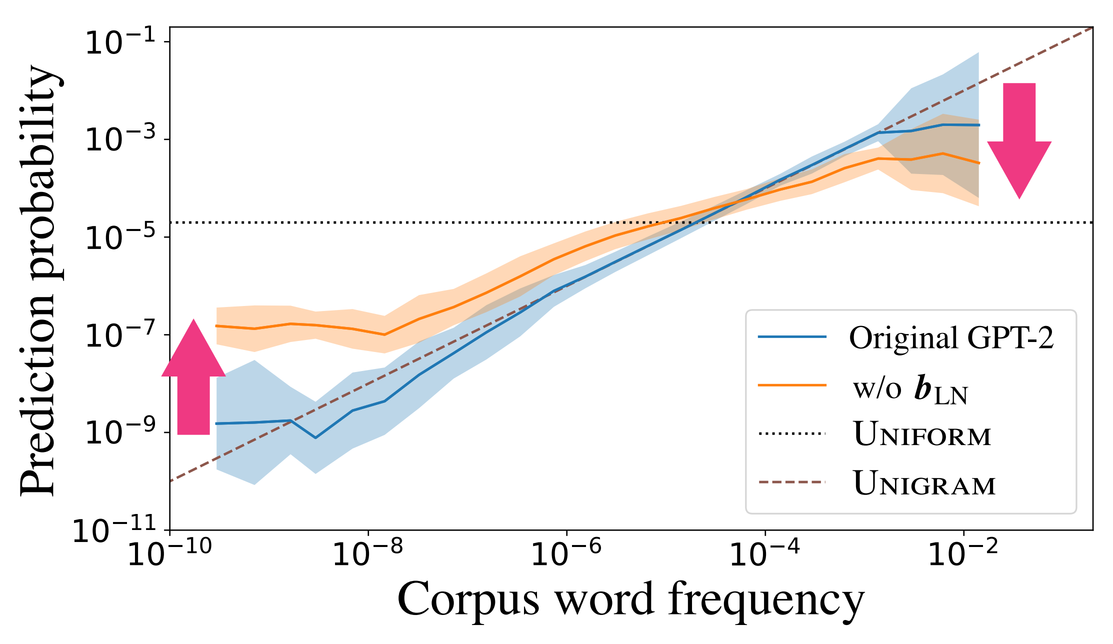
Finding 1: Bias adjusts word prediction according to word frequency

- When removing a bias $\boldsymbol{b}_{\text{LN}}$ (— → —)
  - Probability of high-frequency words is decreased ⬇
  - Probability of high-frequency words is increased ⬆

||

- **Bias $\boldsymbol{b}_{\text{LN}}$ adjusts word prediction**
  - **to promote high-frequency words**
  - **to discourage low-frequency words**

Average prediction distribution by GPT-2 small

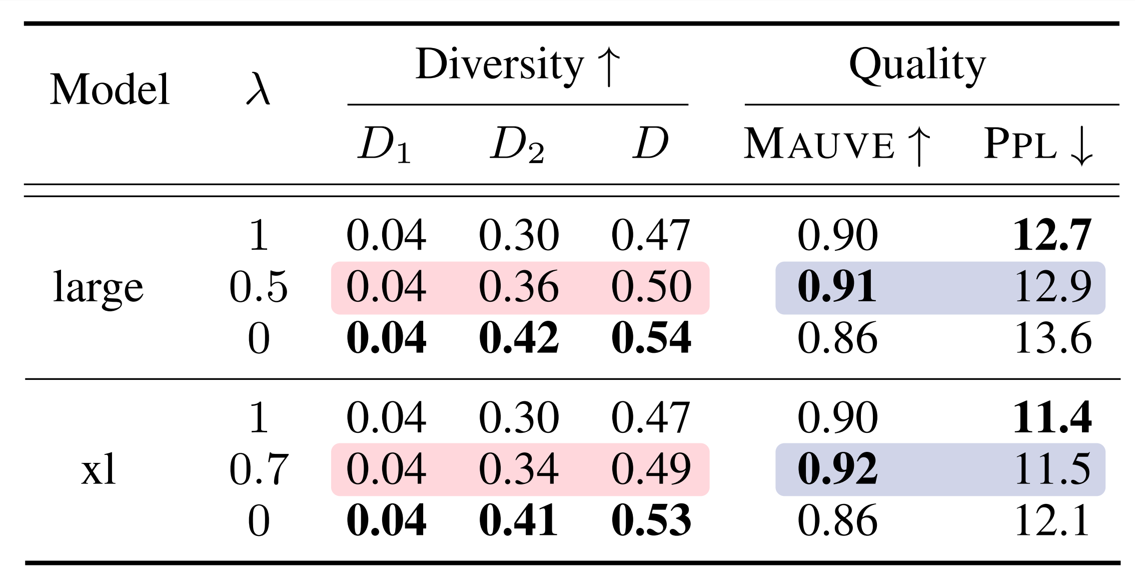


Finding 2: Controlling the bias can encourage more diverse language generation

- Control the bias $\boldsymbol{b}_{\mathrm{LN}}$ with coefficient $\lambda \in [0,1]$

$$\boldsymbol{b}_{\mathrm{LN}} \leftarrow \lambda \boldsymbol{b}_{\mathrm{LN}}$$

- For large models, weakening $\boldsymbol{b}_{\mathrm{LN}}$
  - Improves diversity
  - Maintains quality



Model	λ	Diversity ↑			Quality	
		D_1	D_2	D	MAUVE ↑	PPL ↓
large	1	0.04	0.30	0.47	0.90	**12.7**
	0.5	0.04	0.36	0.50	**0.91**	12.9
	0	**0.04**	**0.42**	**0.54**	0.86	13.6
xl	1	0.04	0.30	0.47	0.90	**11.4**
	0.7	0.04	0.34	0.49	**0.92**	11.5
	0	**0.04**	**0.41**	**0.53**	0.86	12.1

Thank you for listening!

Feel free to ask or comment!

We hope you read the paper
for more details, other findings, and discussions!