LM-Steer: Word Embeddings Are Steers for Language Models

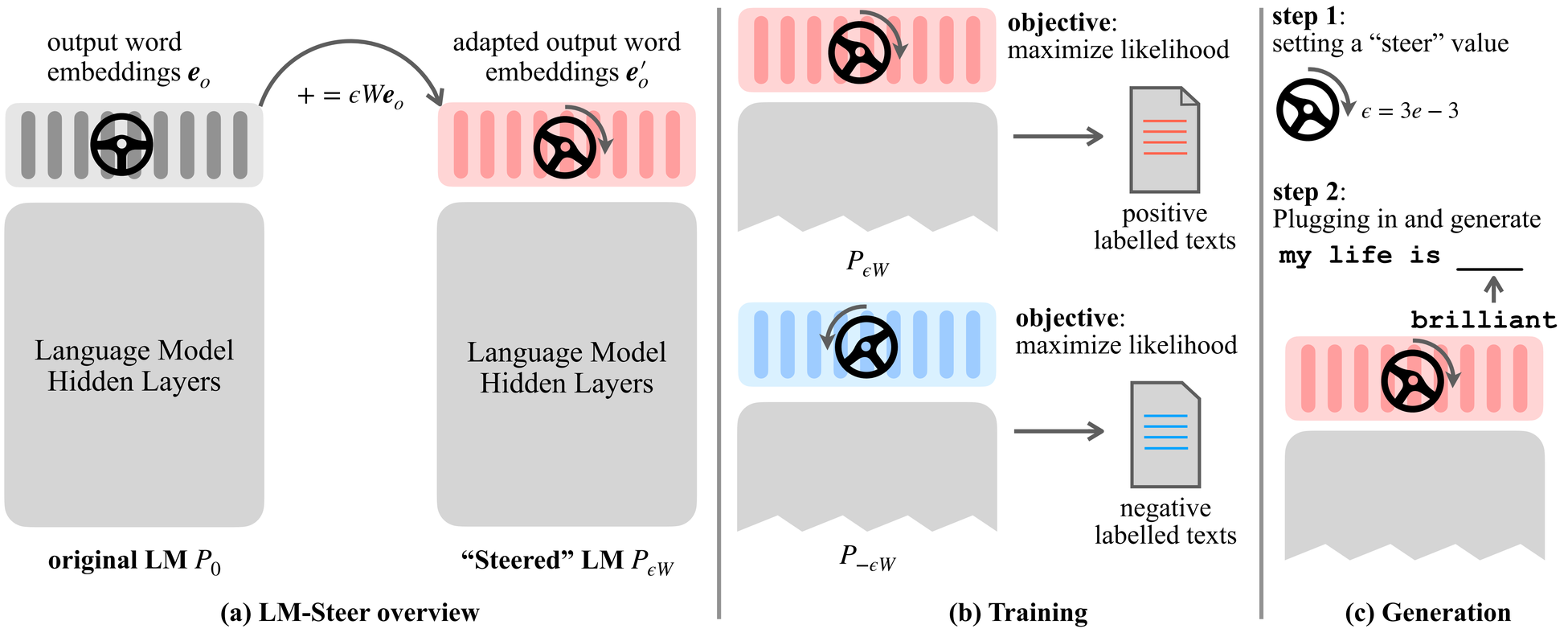
Chi Han, Jialiang Xu, Manling Li,
Yi Fung, Chenkai Sun, Nan Jiang,
Tarek Abdelzaher, Heng Ji

ILLINOIS CS STANFORD

paper codes author

Motivating Question:

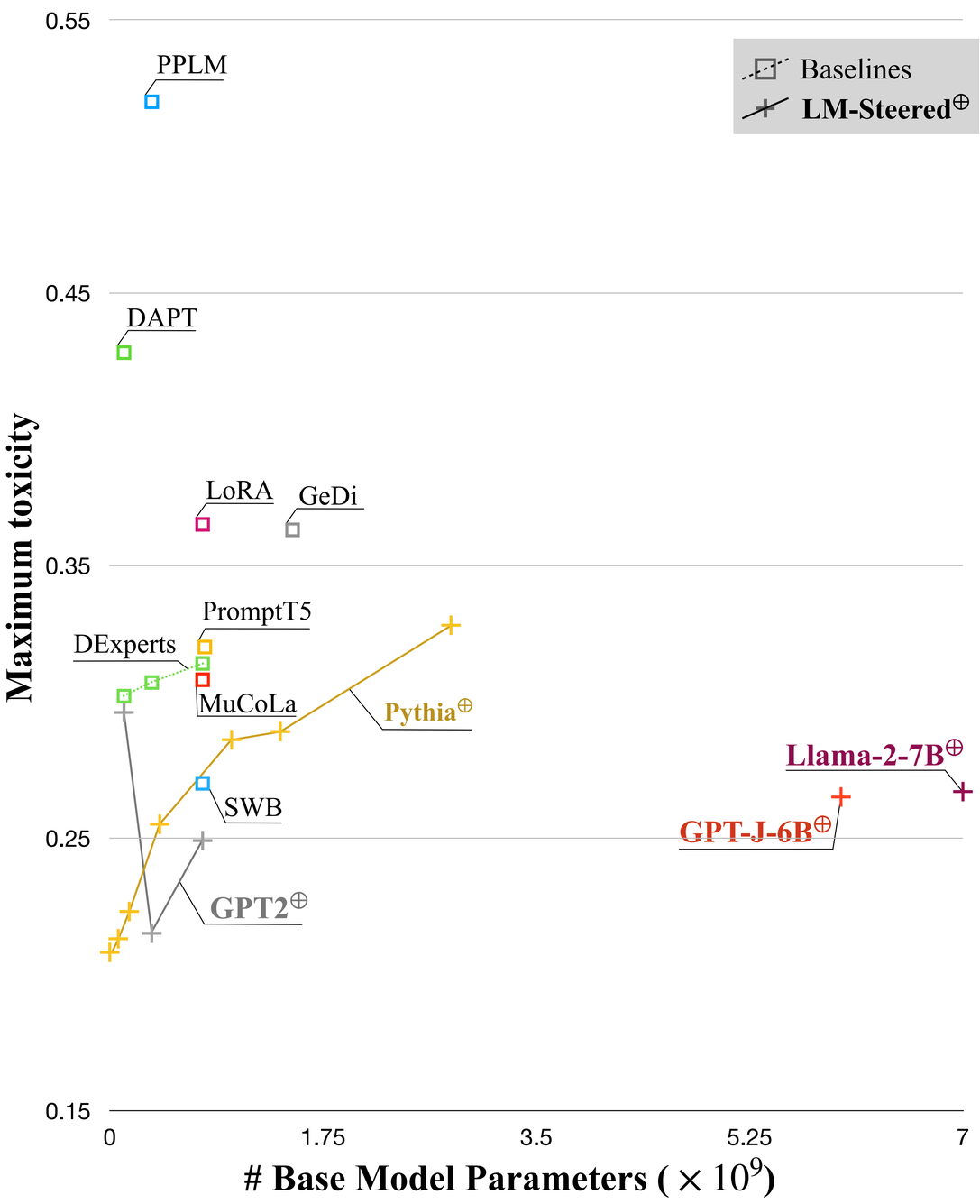
1. <u>What is the role of word embeddings in LLMs?</u>
   The controllers for generations!
2. <u>Can we steer LLMs with this property?</u>
   LM-Steer: a linear transformation in word embeddings
3. <u>What properties does it enjoy?</u>
   Continuous & compositional control, interpreting word embeddings, keyword detection, transferability, etc.



(a) Overview: LM-Steer applies a linear factor $\epsilon W \mathbf{e}_v$ to each word embedding for language model steering

(b) Training: we maximize the likelihood of a positively steered model $P_{\epsilon W}$ on positively labeled texts and vice versa

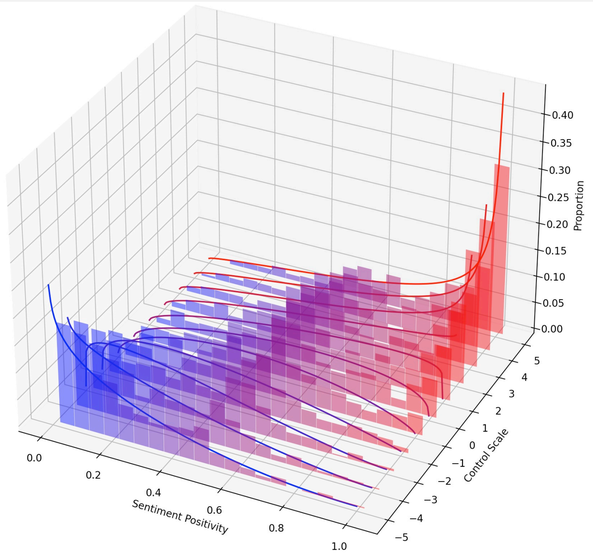
(c) Generation: one can customize a steering value ϵ' and then proceed with normal decoding on the model $P_{\epsilon' W}$



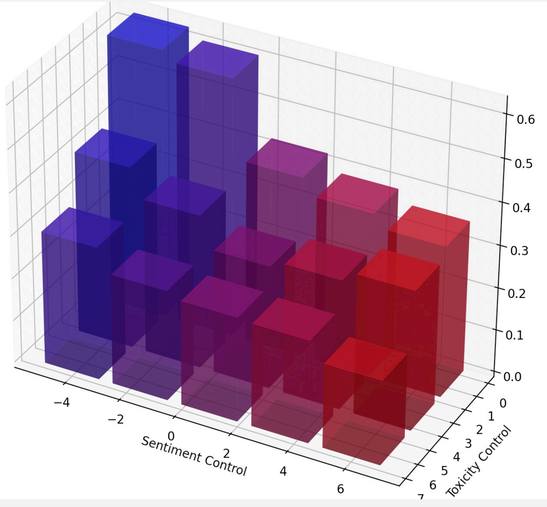
Detoxification task: across base model sizes, LM-Steered (labeled as *LM*⊕ with symbols +) consistently outperforms the other baselines (labeled as *LM* with symbols □).

Dim.	Matched Words
0	mor, bigot, Stupid, retarded, coward, stupid, loser, clown, dumb, Dumb, losers, stupidity, garbage , idiots, fools, idiot, lame
1	stupid, idiot, Stupid, idiots, jerk, pathetic, suck, buff, stupidity, mor, damn, ignorant, fools, dumb , disgusting , damned, narcissistic, troll
3	idiot, godd, damn,
5	Balk, lur, looms, hides, shadows, Whites, slippery, winds
7	bullshit, fiat, shit, lies, injust, manipulation
8	disabled, inactive, whip, emo, partisan, spew, bombed, disconnected, gun, failing, Republicans , defeated, Jeb, blowing , bombard, ineffective, reload, destructive, flo, blown
9	winners, upside

Revealing Toxicity-Associated Word Dimensions with LM-Steer



Continuous sentiment steering



Compositional sentiment and toxicity steering

	LM-Steer	DAPT	GeDi	CTRL	PPLM	DExpert	MuCoLa	LoRA
Parameters	**1.6M**	355M	355M	355M	124M	355M	898M	18M
Speed Ratio	1.24	**1.00**	2.94	3.79	270.11	1.98	24.03	**1.00**

Parameter and Time Efficiency of LM-Steer