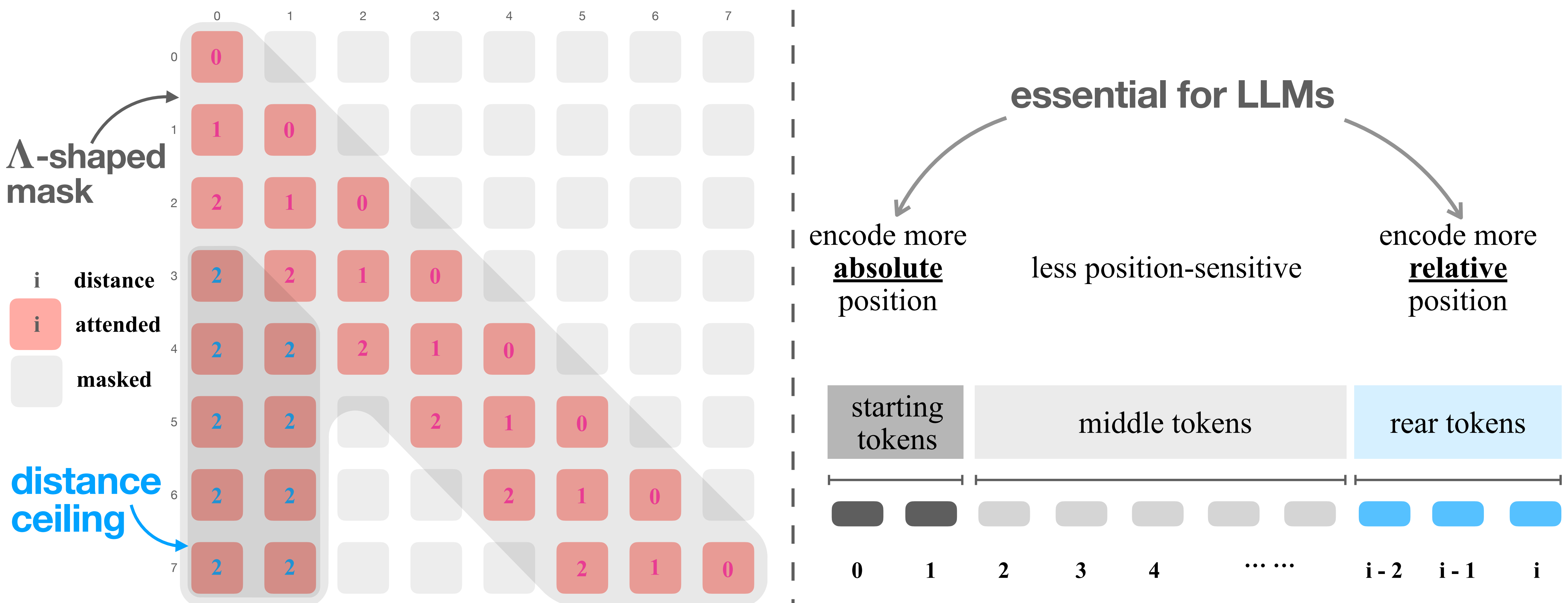
LM-Steer:
Word Embeddings Are Steers for Language Models

Chi Han[1], Jialiang Xu[2], Manling Li[2], Yi Fung[1], Chenkai Sun[1], Nan Jiang[1], Tarek Abdelzaher[1], Heng Ji[1]
[1]UIUC, [2]Stanford

ACL 2024, Outstanding Paper Award, https://arxiv.org/abs/2305.12798
First Author: https://glaciohound.github.io, *chihan3@illinois.edu*
Code Repo: https://github.com/Glaciohound/LM-Steer

A Companion Piece: LM-Infinite

Zero-Shot Extreme Length Generalization for Large Language Models



(a) Proposed Solution: LM-Infinite

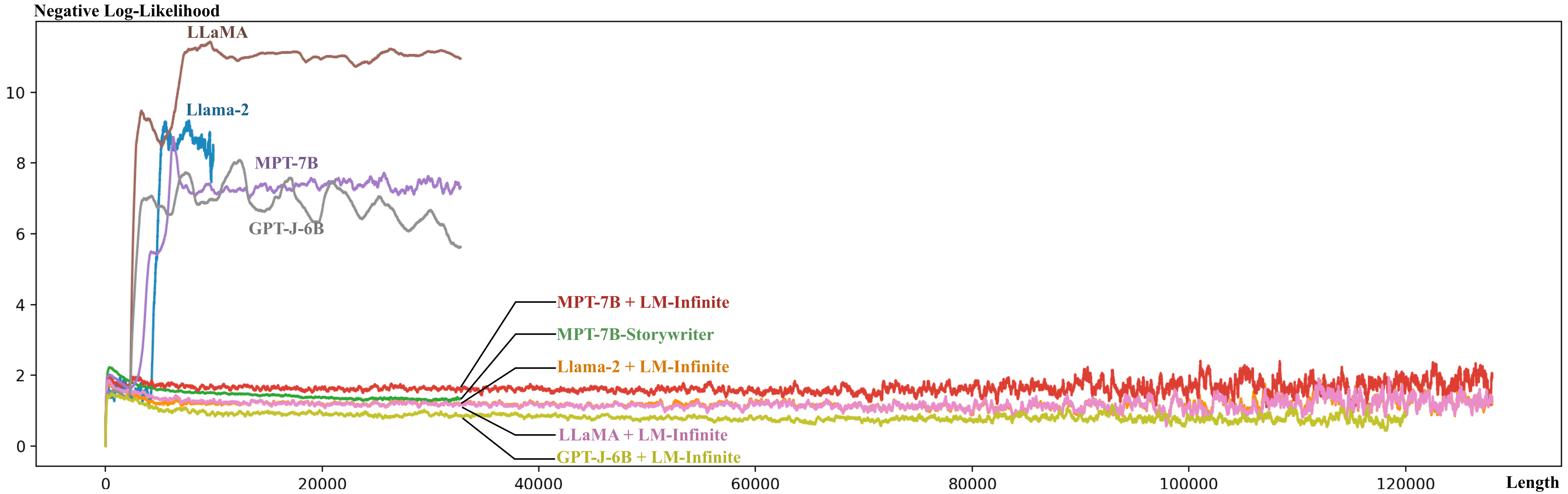
(b) A Conceptual Model of Relative Positional Attention

- Studies the OOD issues in length representation of LMs
- Provides a conceptual model of length representation

NAACL 2024, Outstanding Paper, https://arxiv.org/abs/2308.16137

A Companion Piece: LM-Infinite

Zero-Shot Extreme Length Generalization for Large Language Models

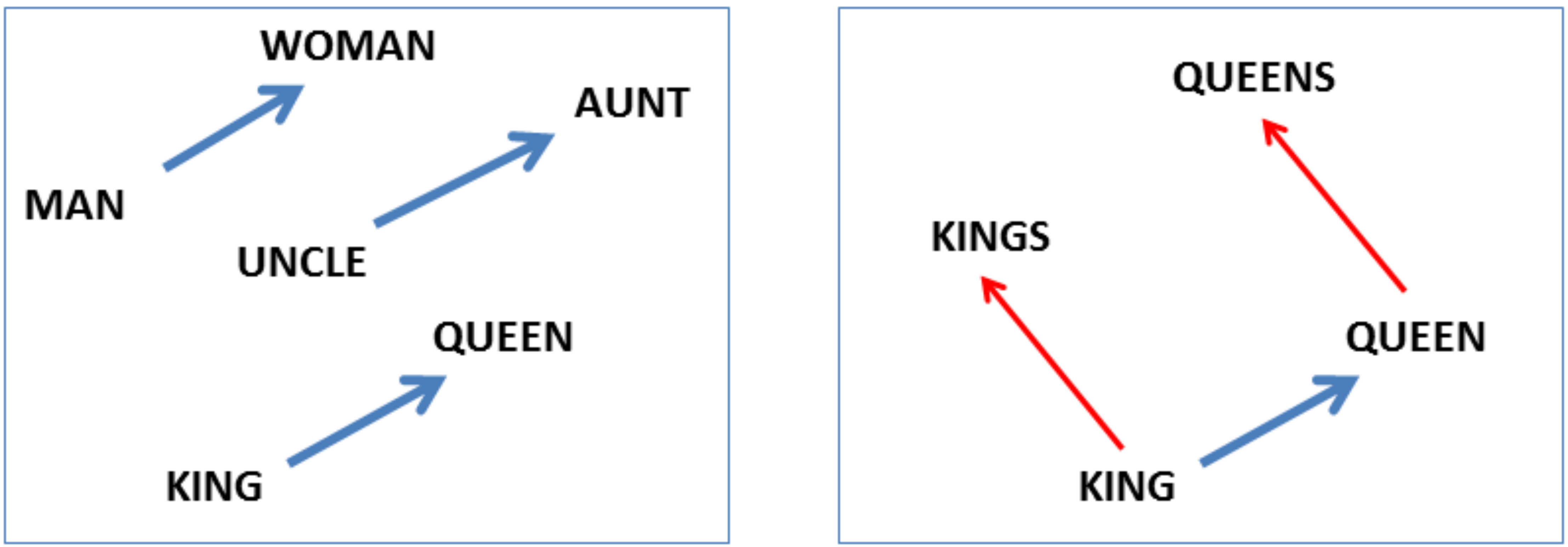


- applies to various modern LLMs without parameter updates
- Extreme generalization to 200M, with downstream task improvements

NAACL 2024, Outstanding Paper, https://arxiv.org/abs/2308.16137

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations

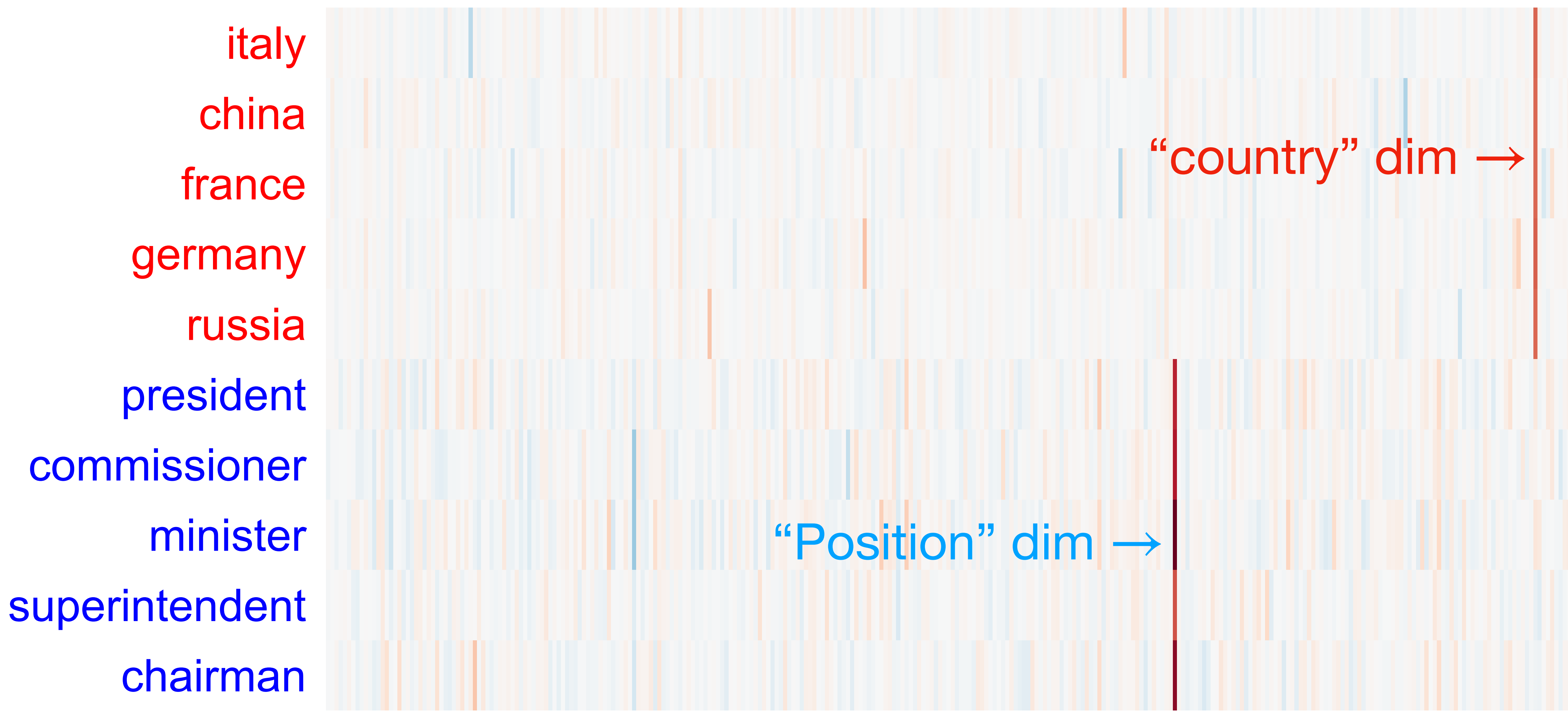


(a) Analogical Relations (metric space)

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013.
Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations

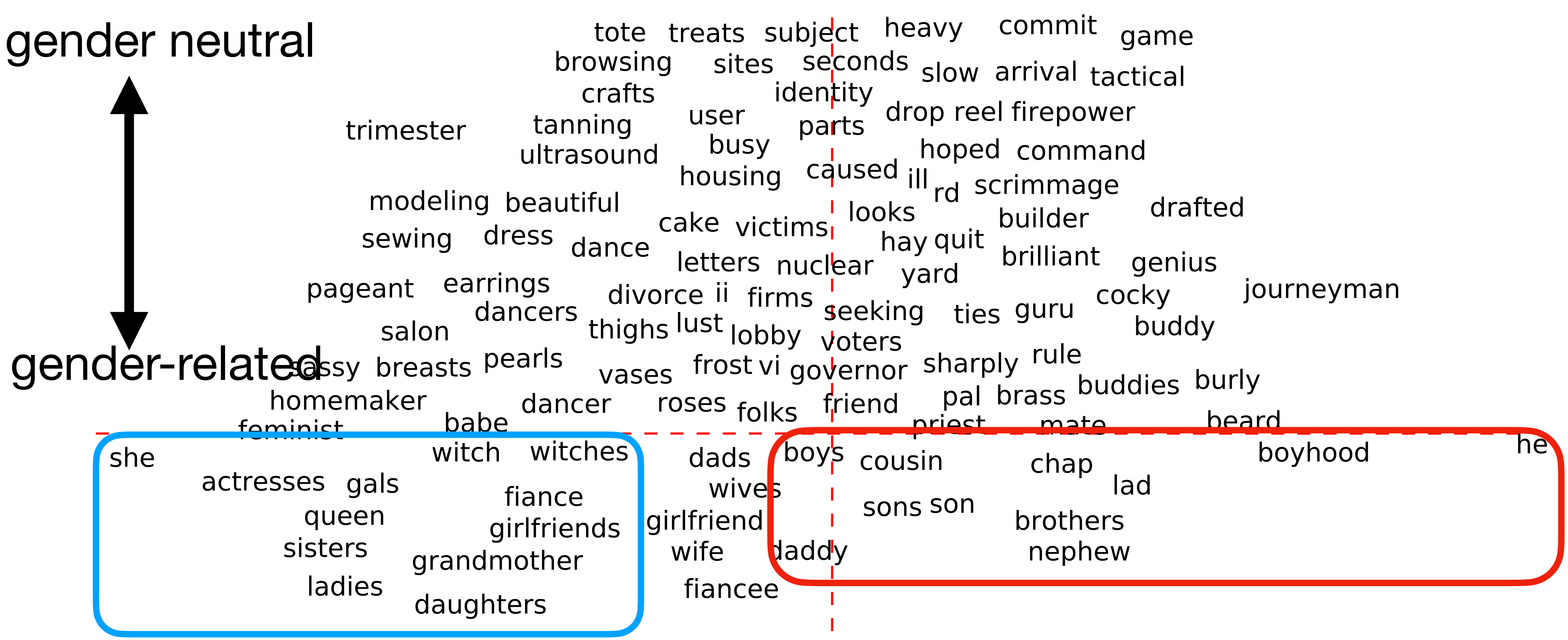


(b) Meaningful Dimensions (linear Space)

Park, Sungjoon, JinYeong Bak, and Alice Oh. "Rotated word vector representations and their interpretability." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

What Do Word Embeddings Embed?

Previous papers mostly focus on word-level interpretations



(b) Meaningful Dimensions (linear Space)

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

What Do Word Embeddings Embed?

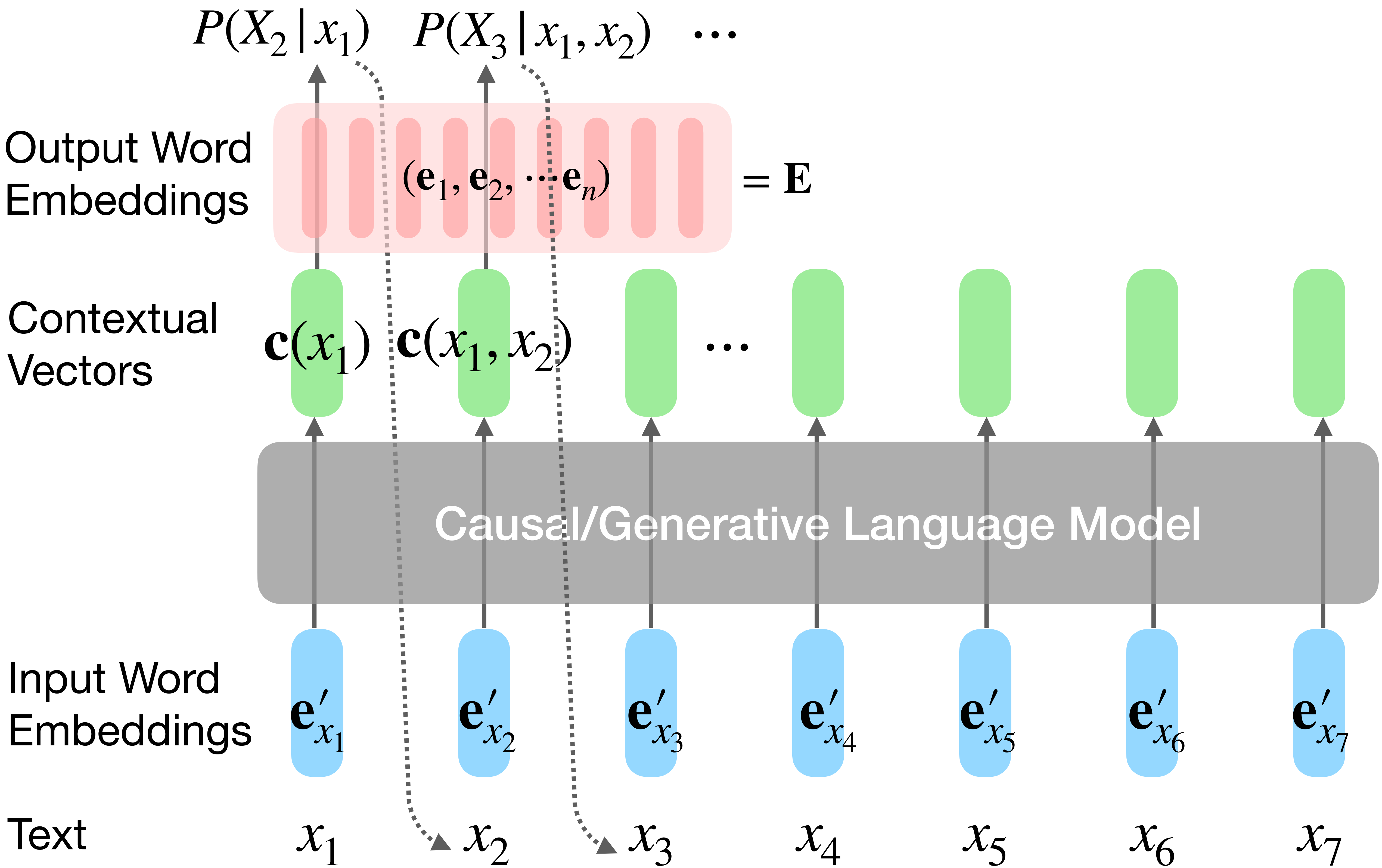
Previous papers mostly focus on word-level interpretations

$\mathbf{u}^1$	$\mathbf{u}^4$	$\mathbf{u}^7$	$\mathbf{u}^8$	$\mathbf{u}^{14}$	$\mathbf{u}^{121}$
lastly	molly	determinants	shyam	famille	jays
outset	sally	biochemical	sanjeev	vrier	strikeouts
ostensibly	toby	intrinsic	meera	autour	halladay
curiously	maggie	qualitative	anupama	naissance	hitters
actuality	valentine	elucidated	deepa	rique	buehrle
crucially	jenny	analytical	rajkumar	diteur	batters
theirs	tracy	psychological	manju	octobre	pitching
importantly	lucy	unger	uday	chambre	phillies
thankfully	carrie	ehrlich	chitra	lettre	rbis
regrettably	elliot	quantitative	vinod	campagne	astros
ironically	susie	integrative	archana	jeune	diamondbacks
aforementioned	laurie	extrinsic	bhanu	jours	homers
paradoxically	cooper	nagel	santosh	septembre	hitless
oftentimes	jill	methodologies	rajesh	enfance	orioles
doubtless	kitty	exogenous	ashok	plon	podsednik
unsurprisingly	charlie	underneath	munna	affaire	baserunners
connelly	shirley	translational	suman	cembre	hitter
merrick	hannah	kuhn	komal	royaume	sox
invariably	annie	functional	subhash	propos	pettitte
dunning	elaine	schweitzer	usha	juin	vizquel
Transition	First Names	Science	Indian Names	French	Baseball

(b) Meaningful Dimensions (linear Space)

Shin, J., Madotto, A., & Fung, P. (2018). Interpreting word embeddings with eigenvector analysis. In 32nd Conference on Neural Information Processing Systems (NIPS 2018), IRASL workshop (pp. 73-81).

Word Embeddings in Causal LMs



Revisit the Question

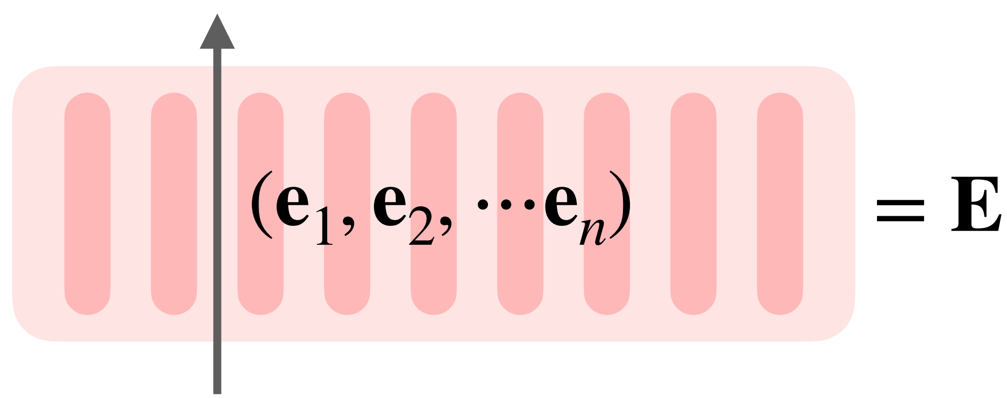
What Do Word Embeddings Embed in LMs?

- LM's optimization objective: generation, alignment, etc.
- LMs learn word embeddings incidentally.
  - But by no means randomly!
- What is the role of word embeddings?

Output Word Embeddings

Projecting to Logits

$$P(X_i \,|\, x_1, \cdots, x_{i-1})$$



$$\mathbf{c}(x_1, \cdots, x_{i-1})$$

$$P(v|\mathbf{c}) = \frac{\exp(\mathbf{c}^\top \mathbf{e}_v)}{\sum_{u \in \mathcal{V}} \exp(\mathbf{c}^\top \mathbf{e}_u)}$$

Output Word Embeddings

A similarity measure

$$logit(\mathbf{c}, \mathbf{e}) = \mathbf{c}^\top \mathbf{e} \;\; : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

- An inner-product space
- $\mathbf{c}, \mathbf{e}$ resides in the same vector space of V
- the direction of $\mathbf{c}$: relatedness direction
- the length of $\mathbf{c}$: how concentrated the distribution is

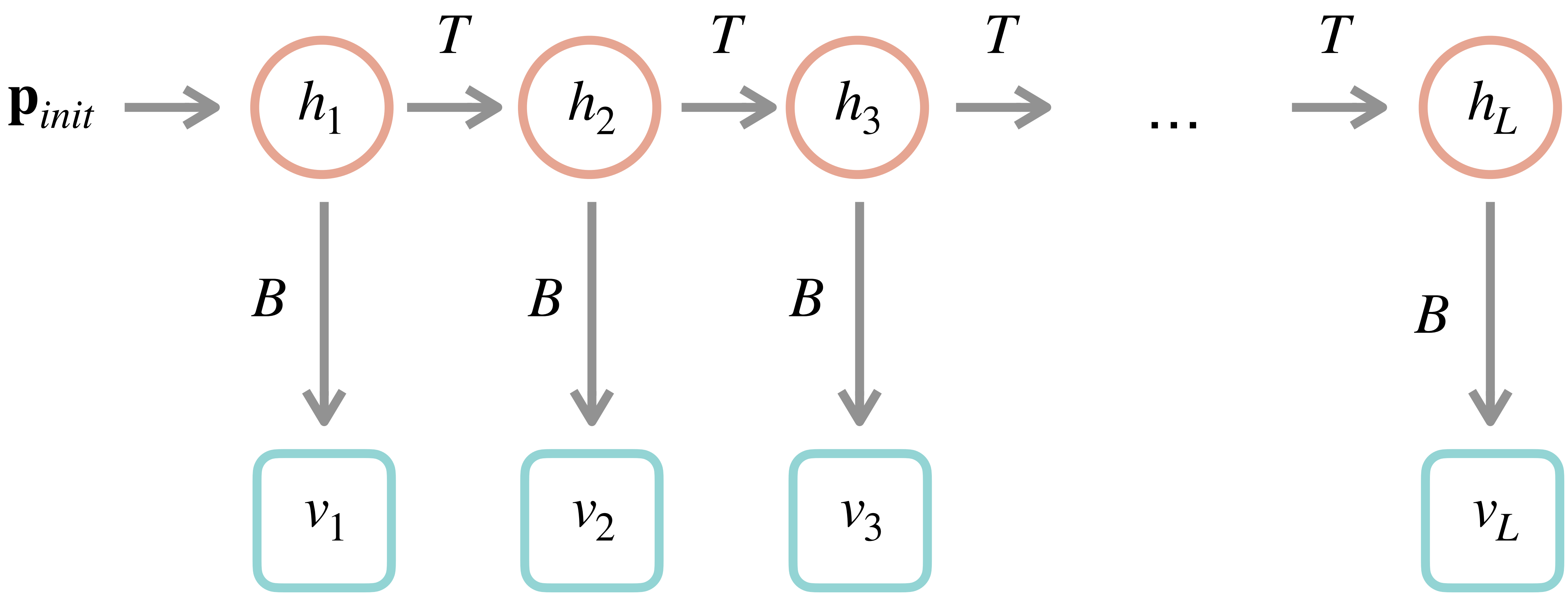
Output Word Embeddings

A Dimension Reduction

$$\mathbf{E} : [1..n] \rightarrow \mathbb{R}^d$$

- when $k = |\mathscr{V}|$ can theoretically express any distribution
- when $k < |\mathscr{V}|$, compresses (embeds) words so they are inter-related
    - but, in what way?

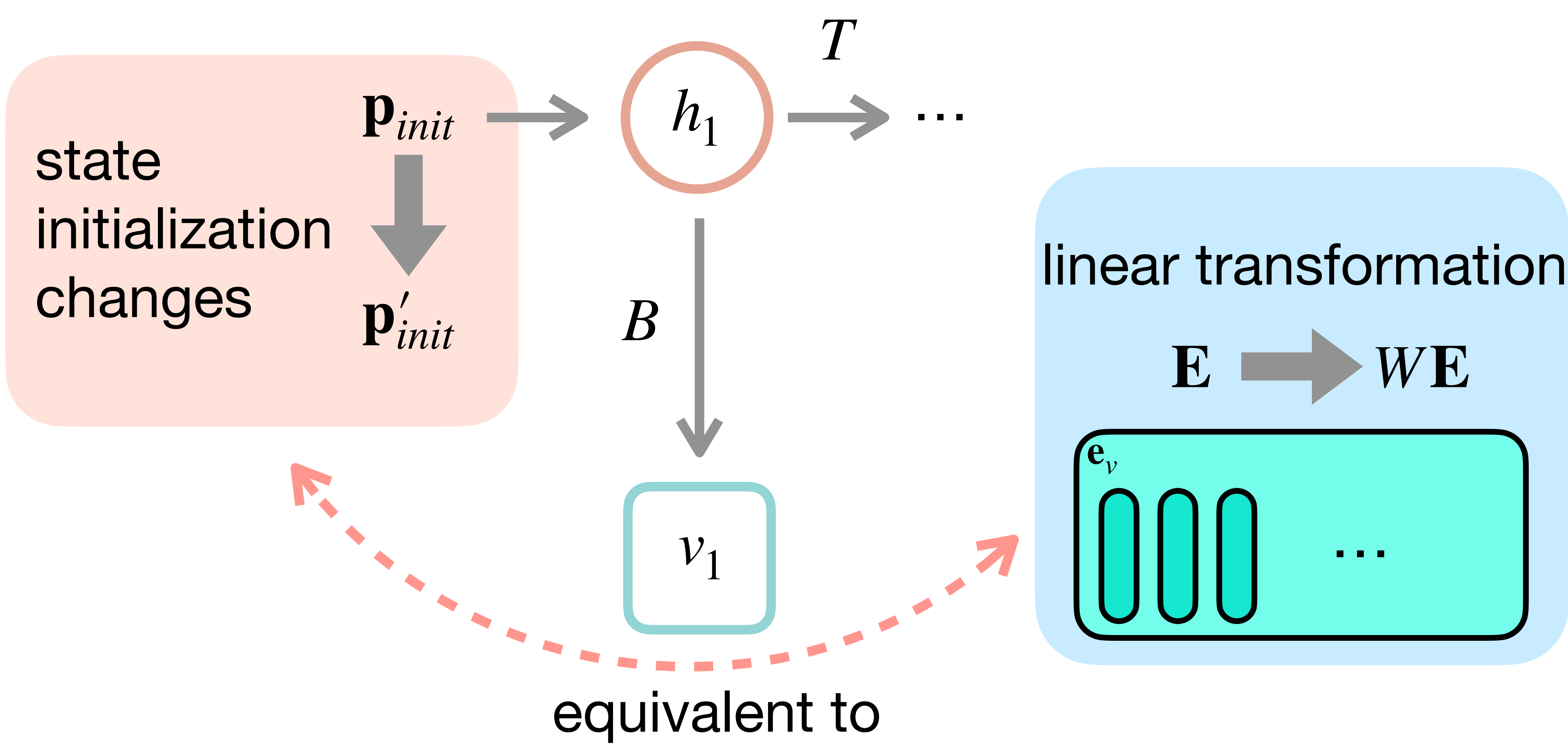
HMM as A Theoretical Framework



$$P_{HMM}(v_1, \cdots, v_L; \mathbf{p}_{init}) = \mathbf{p}_{init}^{\top} T \left(\prod_{i=1}^{L-1} diag(\mathbf{p}(v_i))T \right) \mathbf{p}(v_L)$$

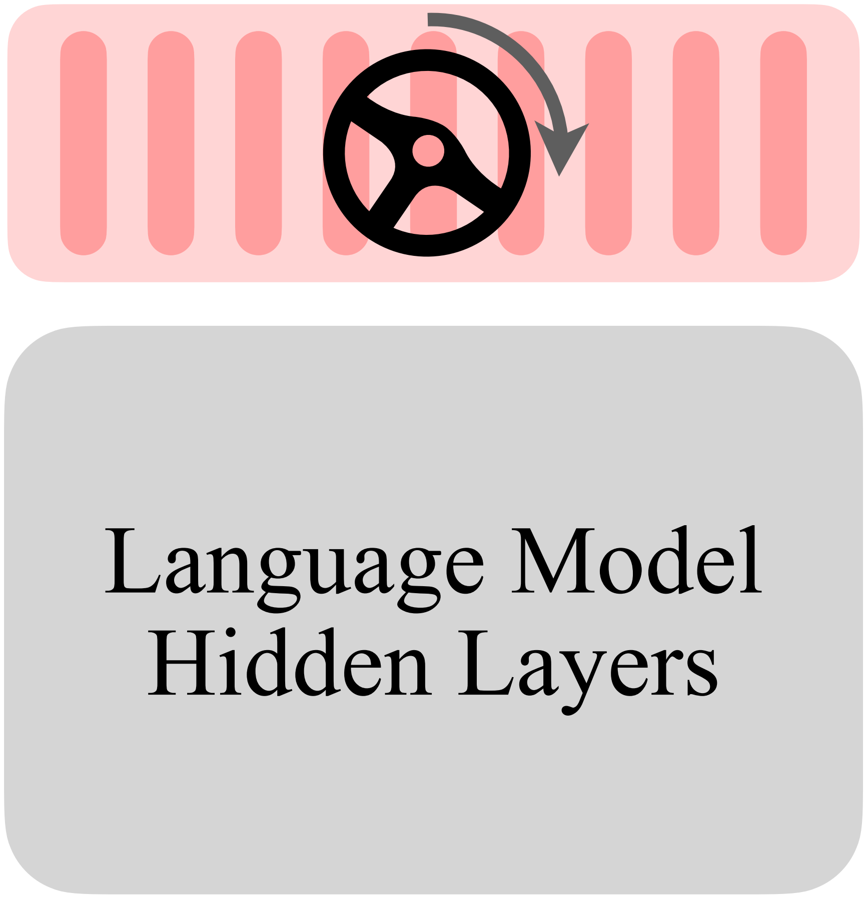
Sequence Shift ≈ Word Embedding Transform

- **Theorem (Informal)**: steering between text distribution is associated with a linear transformation on word embedding space under assumptions.



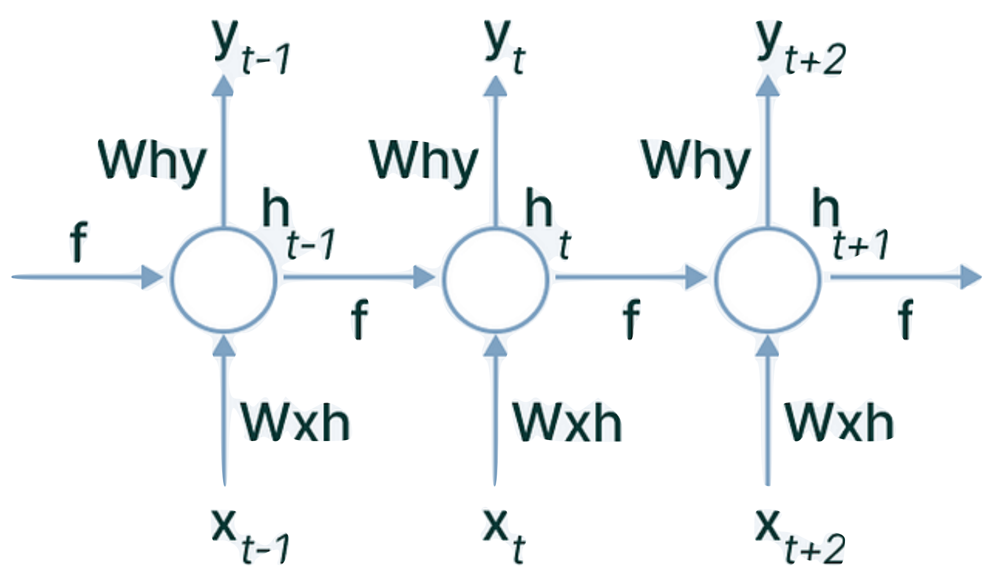
Word Embeddings Are Steers

An Intuitive Explanation

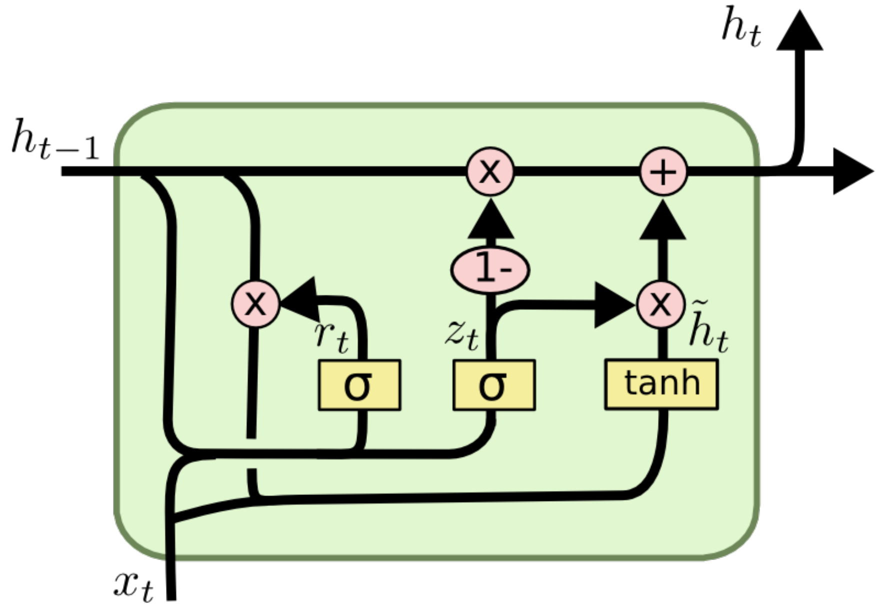


- Non-trivial claim as it connects word distributions and sequence distributions

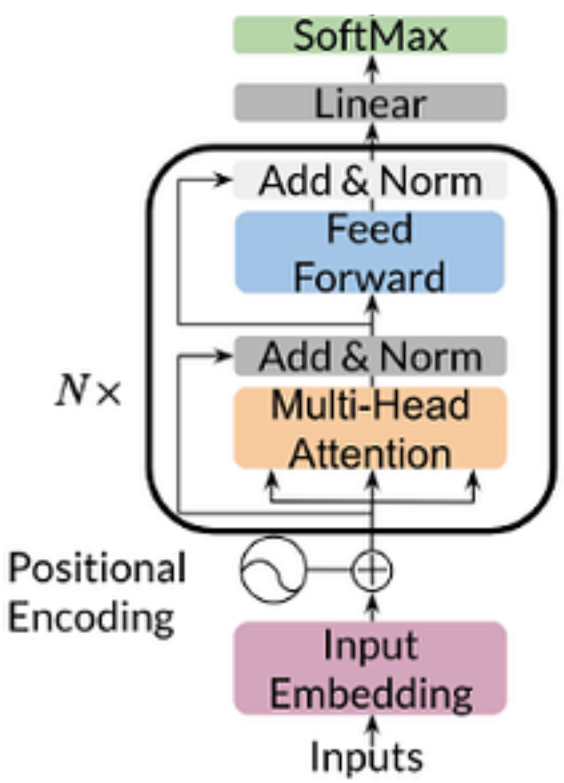
Theoretical Generality



RNNs



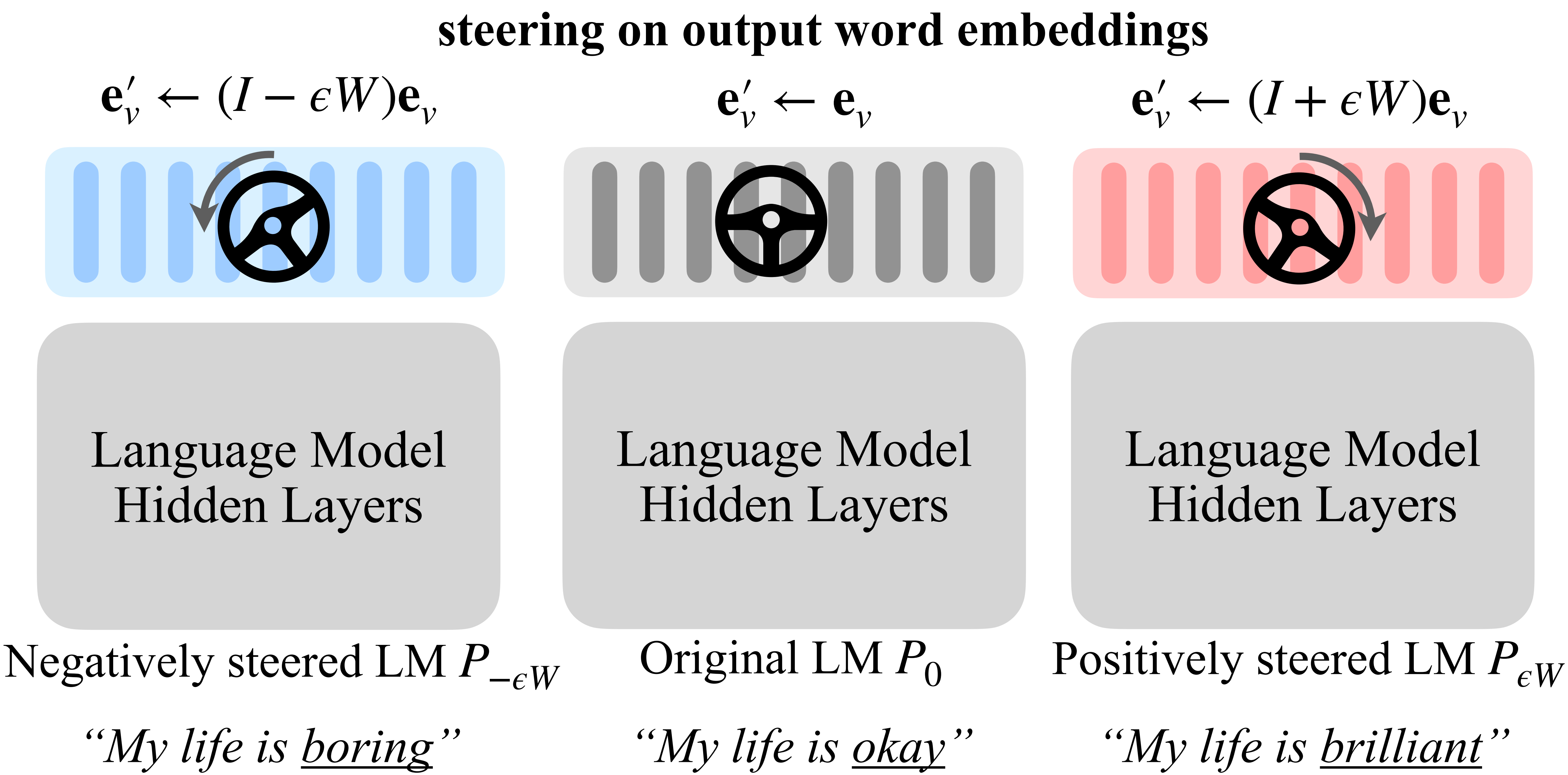
LSTMs



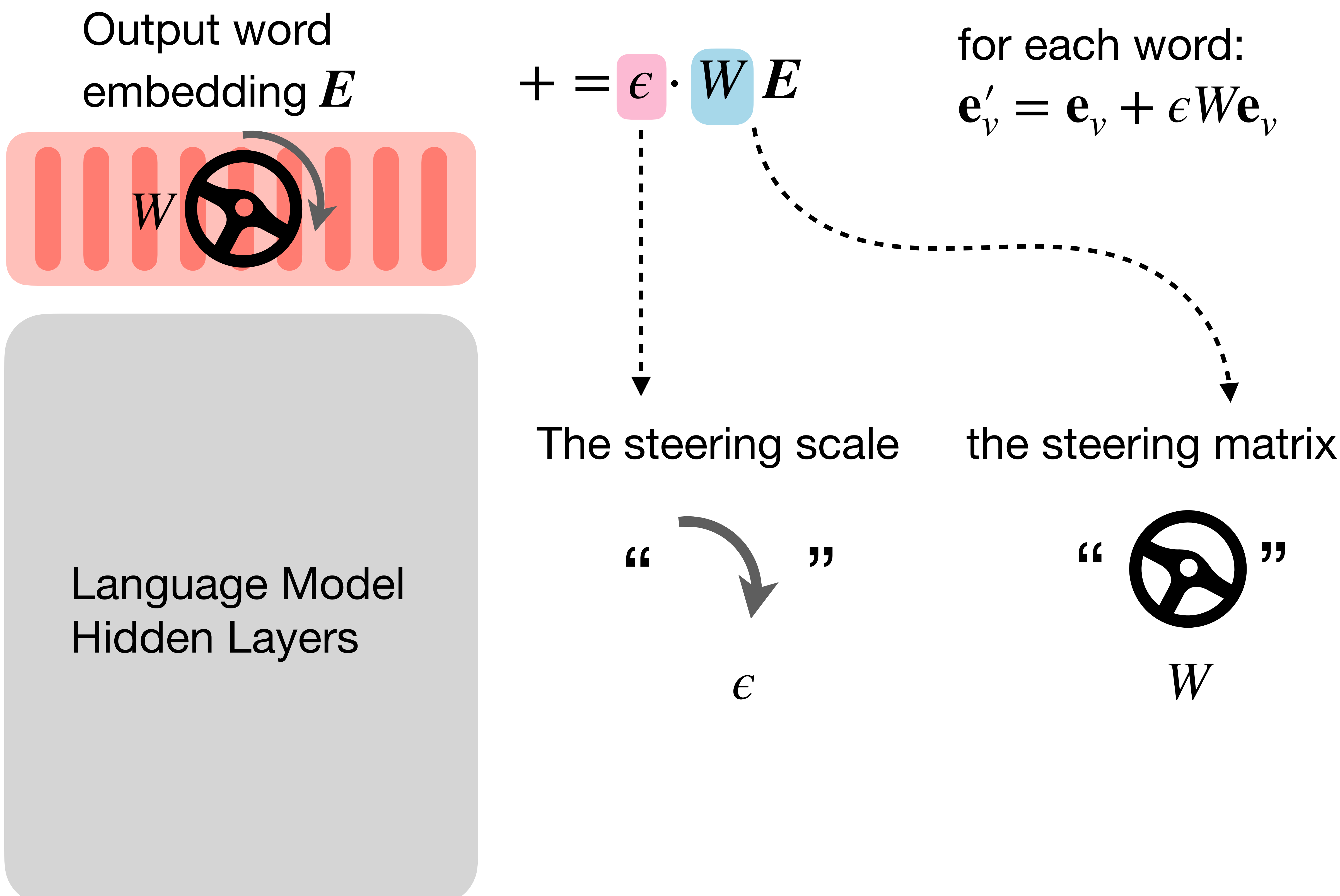
Transformers

https://miro.medium.com/v2/resize:fit:1200/1*5NhjY5OH8HKpi5oHuEMxTg.png
https://colah.github.io/posts/2015-08-Understanding-LSTMs/img/LSTM3-var-GRU.png

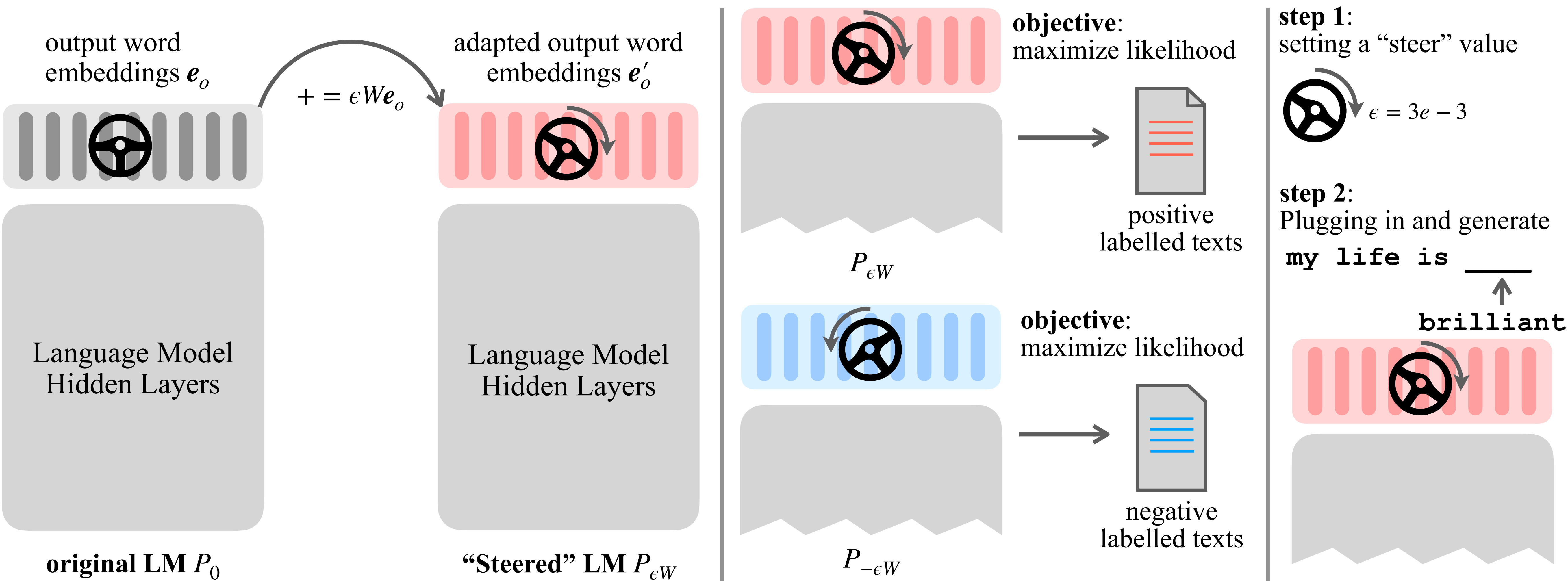
LM-Steer


steering on output word embeddings
$\mathbf{e}'_v \leftarrow (I - \epsilon W)\mathbf{e}_v$
$\mathbf{e}'_v \leftarrow \mathbf{e}_v$
$\mathbf{e}'_v \leftarrow (I + \epsilon W)\mathbf{e}_v$
Language Model Hidden Layers
Language Model Hidden Layers
Language Model Hidden Layers
Negatively steered LM $P_{-\epsilon W}$
Original LM P_0
Positively steered LM $P_{\epsilon W}$
"My life is boring"
"My life is okay"
"My life is brilliant"

LM-Steer Broken Down



Training & Inference



(a) LM-Steer overview **(b) Training** **(c) Generation**

Detoxification

Main metric

	Model	Backbone Size	Toxicity↓ Max. toxicity	Toxicity↓ Toxicity prob.	Fluency Output ppl.↓	Diversity↑ Dist-1	Diversity↑ Dist-2	Diversity↑ Dist-3
	GPT-2 (original)	117M	0.527	0.520	25.45	0.58	0.85	0.85
optimization-based	PPLM (10%)	345M	0.520	0.518	32.58	0.58	0.86	0.86
fine-tuning	DAPT	117M	0.428	0.360	31.21	0.57	0.84	0.84
conditioned generation	GeDi	1.5B	0.363	0.217	60.03	0.62	0.84	0.83
offseting logits	$\text{DExperts}_{\text{base}}$	117M	0.302	0.118	38.20	0.56	0.82	0.83
offseting logits	$\text{DExperts}_{\text{medium}}$	345M	0.307	0.125	32.51	0.57	0.84	0.84
offseting logits	$\text{DExperts}_{\text{large}}$	762M	0.314	0.128	32.41	0.58	0.84	0.84
prompting	PromptT5	780M	0.320	0.172	55.1	0.58	0.76	0.70
optimization-based	MuCoLa	762M	0.308	0.088	29.92	0.55	0.82	0.83
efficient finetuning	LoRA	762M	0.365	0.210	21.11	0.53	0.85	0.86
word blacklist	Soft-Blacklist	762M	0.270	0.154	18.28	0.53	0.81	0.83
our model	$\text{LM-Steer}_{\text{base}}$	117M	$0.296_{\pm 0.018}$	$0.129_{\pm 0.012}$	36.87	0.54	0.86	0.86
our model	$\text{LM-Steer}_{\text{medium}}$	345M	$\mathbf{0.215}_{\pm 0.015}$	$\mathbf{0.059}_{\pm 0.029}$	43.56	0.56	0.83	0.84
our model	$\text{LM-Steer}_{\text{large}}$	762M	$0.249_{\pm 0.007}$	$0.089_{\pm 0.009}$	28.26	0.55	0.84	0.84

LM-Steer outperforms each baseline under similar **model sizes**

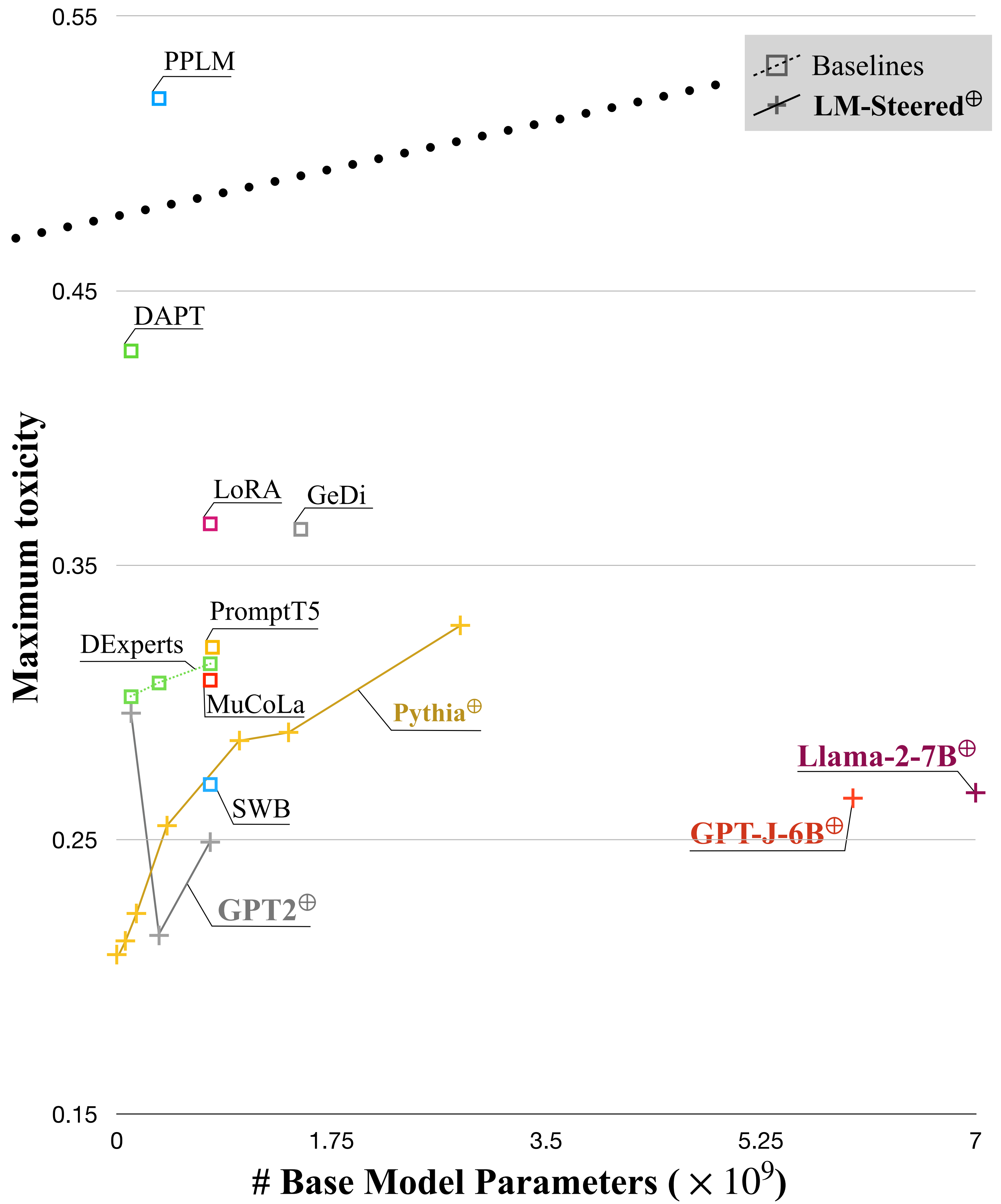
Detoxification

Holistic Comparison

Baselines

LM-Steered⊕

- Across base model sizes, LM-Steered *GPT2 family, Pythia family, GPT-J* and *Llama-2-7B* models (+) consistently outperform other baselines (□) on detoxification.



Detoxification

Pairwise Human Evaluation

	LM-Switch	Tie	LoRA	LM-Switch	Tie	GPT-2	LM-Switch	Tie	DExperts
Detoxified	**19.0**	69.5	11.5	**24.5**	56.5	19.0	**24.0**	56.5	19.5
Fluent	**21.0**	69.0	10.0	21.0	57.5	**21.5**	**25.0**	52.0	23.0
Topical	**18.0**	69.5	12.5	**32.0**	47.0	21.0	**32.0**	56.5	11.5

Metrics

Detoxification

Pairwise Human Evaluation

Baselines: Parameter efficient tuning | Original Model | Controlled generation

	LM-Switch	**Tie**	**LoRA**	**LM-Switch**	**Tie**	**GPT-2**	**LM-Switch**	**Tie**	**DExperts**
Detoxified	**19.0**	69.5	11.5	**24.5**	56.5	19.0	**24.0**	56.5	19.5
Fluent	**21.0**	69.0	10.0	21.0	57.5	**21.5**	**25.0**	52.0	23.0
Topical	**18.0**	69.5	12.5	**32.0**	47.0	21.0	**32.0**	56.5	11.5

Detoxification

Pairwise Human Evaluation

	LM-Switch	Tie	LoRA	LM-Switch	Tie	GPT-2	LM-Switch	Tie	DExperts
Detoxified	**19.0**	69.5	11.5	**24.5**	56.5	19.0	**24.0**	56.5	19.5
Fluent	**21.0**	69.0	10.0	21.0	57.5	**21.5**	**25.0**	52.0	23.0
Topical	**18.0**	69.5	12.5	**32.0**	47.0	21.0	**32.0**	56.5	11.5

Better than the baselines on 8 out of 9 tracks

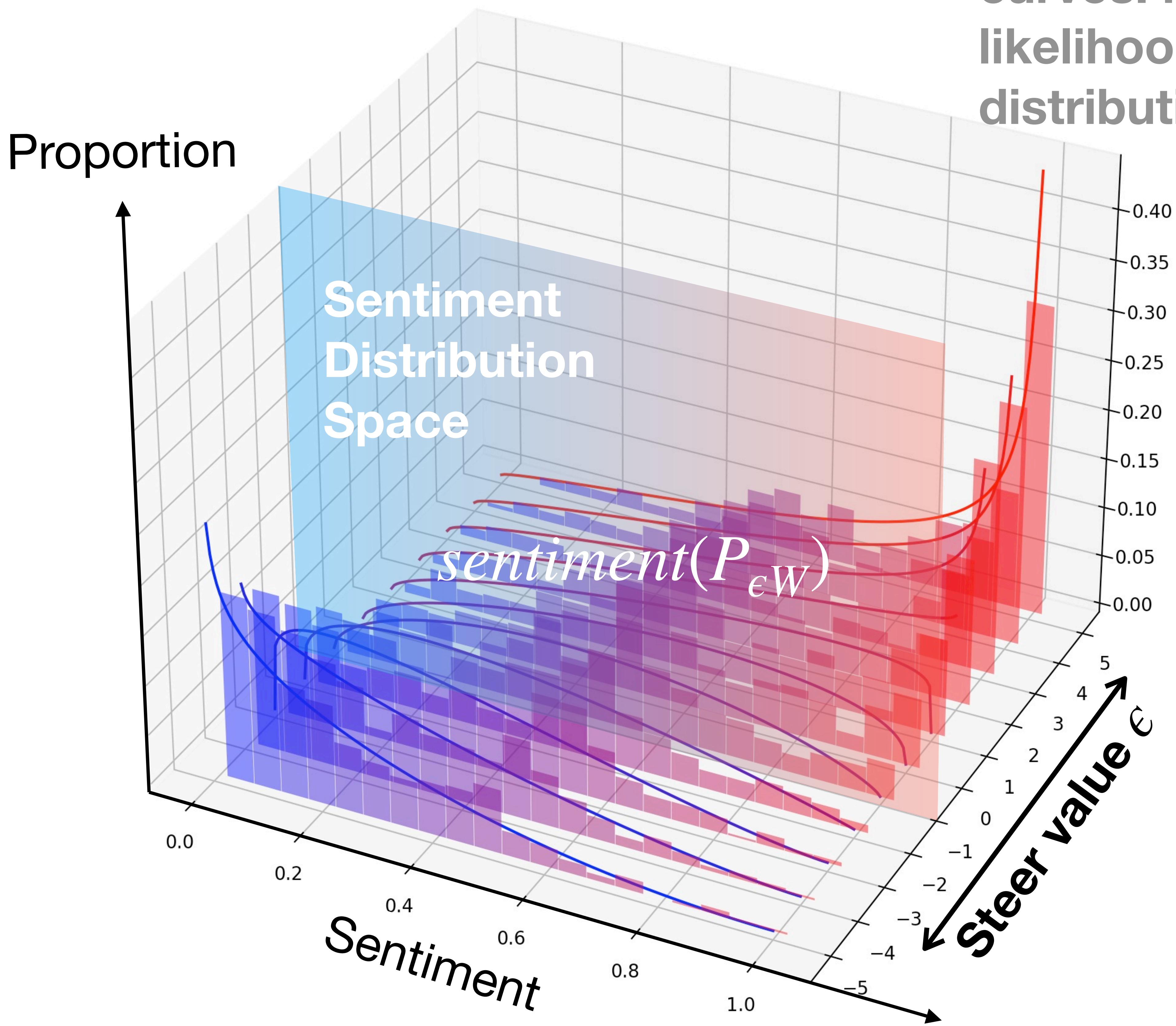
Sentiment Control

- Despite being simpler and smaller
- LM-Steer gets the 1st metrics on the positive sentiment and 2nd to 3rd place on the negative sentiment.

Target	Model	Sentiment Positivity / %			Fluency	Diversity↑		
		Positive prompts	Neutral prompts	Negative prompts	Output ppl.↓	Dist-1	Dist-2	Dist-3
Positive↑	LM-Steer$_{large}$		90.70	41.23	41.20	0.46	0.78	0.83
	LM-Steer$_{medium}$		**95.36**	56.98	67.68	0.46	0.77	0.80
	LM-Steer$_{base}$		90.46	**57.26**	54.38	0.47	0.78	0.81
	Soft-Blacklist		86.40	25.64	99.46	0.42	0.76	0.81
	LoRA		26.88	7.20	158.56	0.57	0.82	0.83
	DExperts$_{large}$		94.46	36.42	45.83	0.56	0.83	0.83
	DExperts$_{medium}$		94.31	33.20	43.19	0.56	0.83	0.83
	DExperts$_{small}$		94.57	31.64	42.08	0.56	0.83	0.84
	DExperts (pos)		79.83	43.80	64.32	0.59	0.86	0.85
	GeDi		86.01	26.80	58.41	0.57	0.80	0.79
	DAPT		77.24	14.17	30.52	0.56	0.83	0.84
	PPLM (10%)		52.68	8.72	142.11	0.62	0.86	0.85
	PromptT5		68.12	15.41	37.3	0.58	0.78	0.72
	GPT-2 (original)	99.08	50.02	0.00	29.28	0.58	0.84	0.84
Negative↓	PromptT5	69.93	25.78		48.6	0.60	0.78	0.70
	PPLM (10%)	89.74	39.05		181.78	0.63	0.87	0.86
	DAPT	87.43	33.28		32.86	0.58	0.85	0.84
	GeDi	39.57	8.73		84.11	0.63	0.84	0.82
	DExperts (neg)	61.67	24.32		65.11	0.60	0.86	0.85
	DExperts$_{small}$	45.25	3.85		39.92	0.59	0.85	0.84
	DExperts$_{medium}$	40.21	3.79		43.47	0.59	0.85	0.84
	DExperts$_{large}$	**35.99**	**3.77**		45.91	0.60	0.84	0.83
	LoRA	57.71	20.08		192.13	0.55	0.78	0.79
	Soft-Blacklist	73.72	14.28		50.95	0.38	0.70	0.76
	LM-Steer$_{base}$	57.26	10.12		51.37	0.49	0.77	0.79
	LM-Steer$_{medium}$	52.32	7.10		71.48	0.47	0.77	0.79
	LM-Steer$_{large}$	54.84	8.02		57.74	0.48	0.78	0.80

Continuous Steering

curves: maximal likelihood beta-distribution



Continuous Steering

Steer	Generation	word toxicity level	# toxic phrases
-5e-3	What **moron** said that **stupid** comment.	“moron”, “stupid”	2
-3e-3	What’s **stupid** is **stupid**, right?	“stupid”	2
-1e-3	What’s this? You think that your religion, your culture, your country are **not good enough**?	“not good enough”	1
0	What’s more, it makes for a fun, cheap, and efficient way to improve the performance of your car engine and to make your driving that much safer.	—	0
1e-3	What’s more, it makes for a fun, cheap, and efficient way to improve the performance of your car engine and motor.	—	0
3e-3	What’s on your mind? What’s on your mind?	—	0
5e-3	What’s on Netflix? If you can’t figure out what’s being watched on Netflix, you need to figure out what are people watching!	—	0

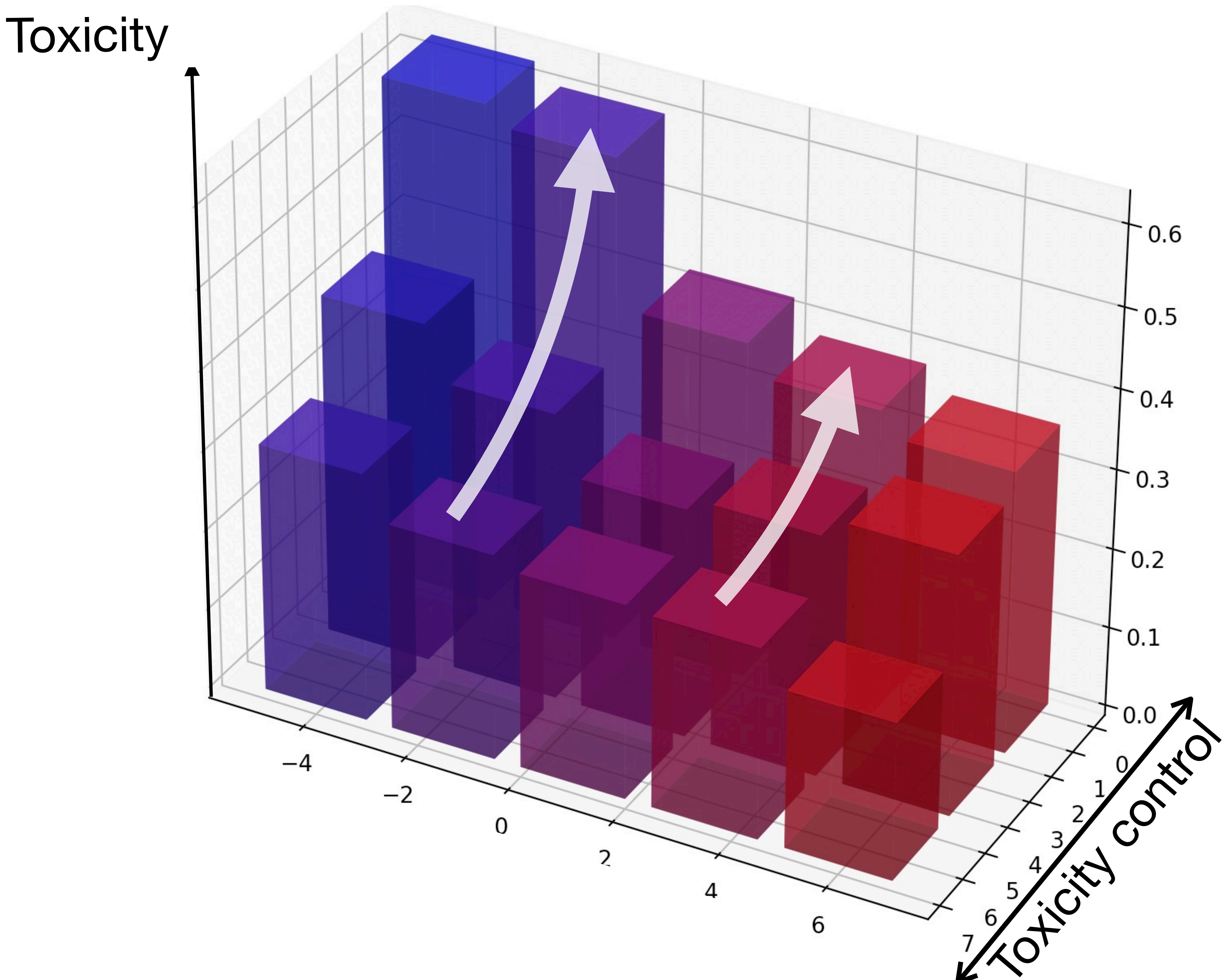
Compositional Steering

LM-Steer 1: $P_{\epsilon_1 W_1}$

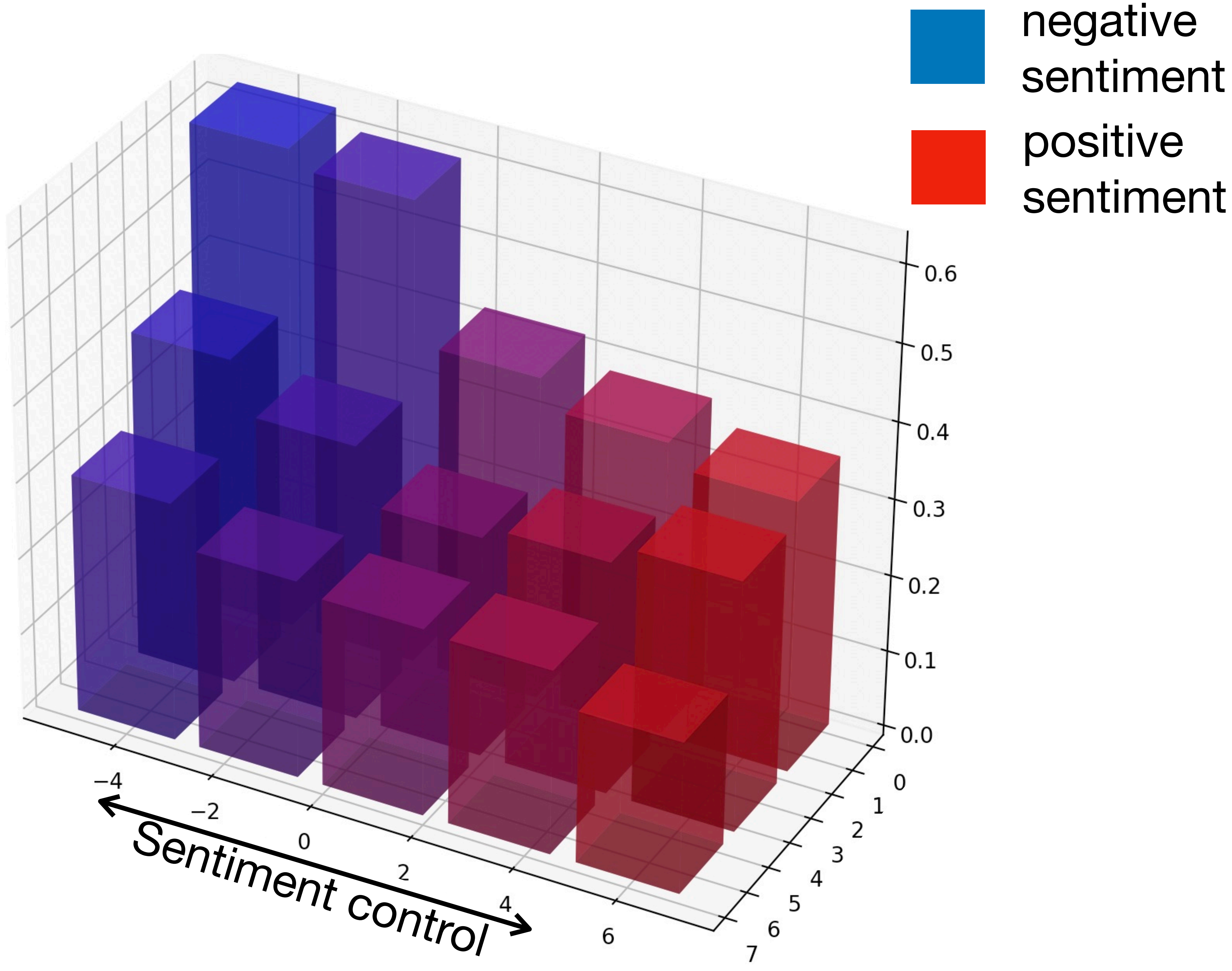
LM-Steer 2: $P_{\epsilon_2 W_2}$

Combined LM-Steer: $P_{\epsilon_1 W_1 + \epsilon_2 W_2}$

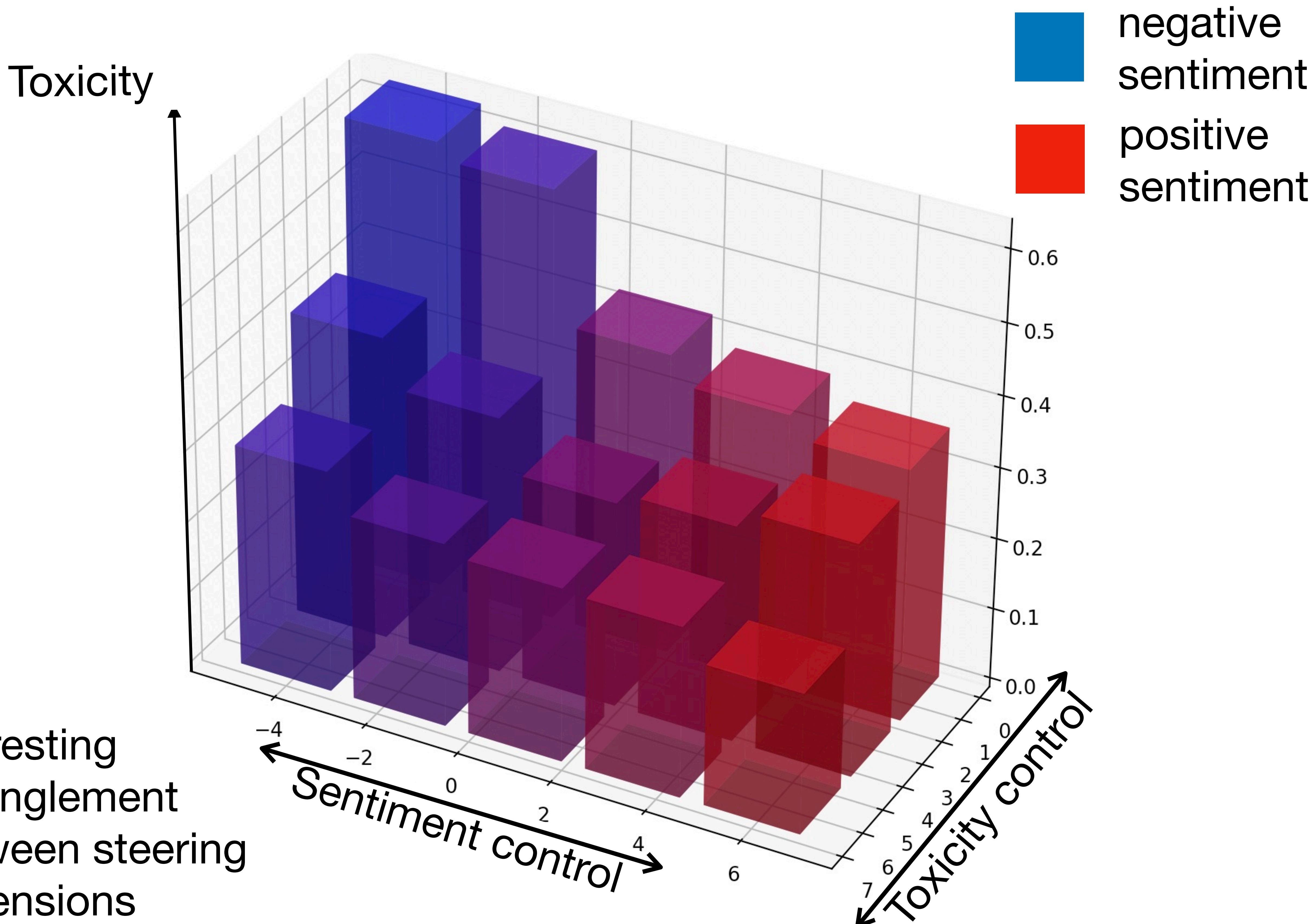
Compositional Steering



Compositional Steering



Compositional Steering



Interesting entanglement between steering dimensions

Transferring to Another LM

LM-Steer defines a bilinear form on the shared space of $\mathbf{c}$ and $\mathbf{e}$

$$\Delta logit(\mathbf{c}, \mathbf{e}) = \epsilon \mathbf{c}^\top W \mathbf{e} \quad : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

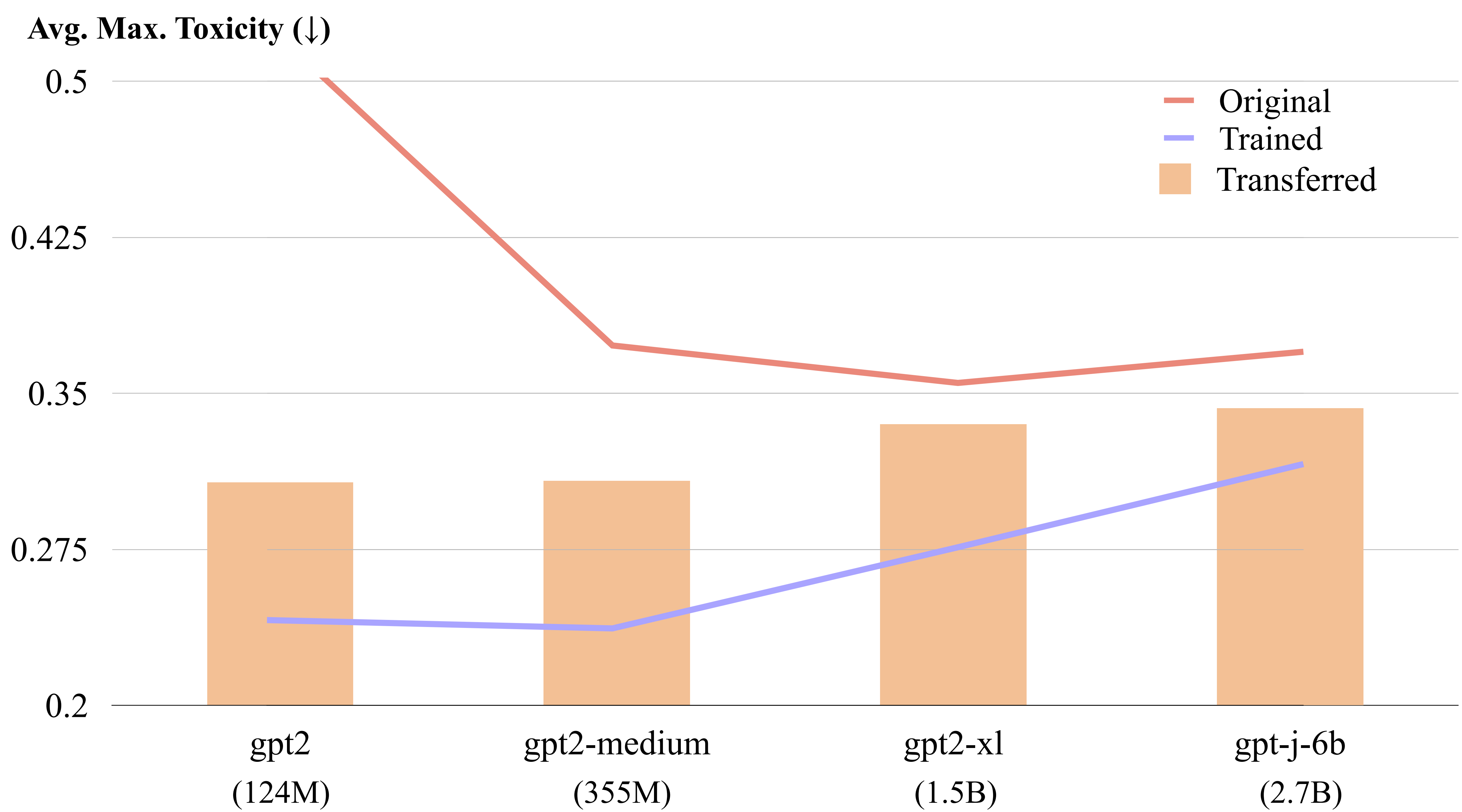
Two transfer to another set of word embeddings: $\mathbf{E} \to \mathbf{E}'$

Assuming an approximate linear transform $\mathbf{E} \approx H\mathbf{E}', \mathbf{c} \approx H\mathbf{c}'$

The equivalent steer term is $\Delta logit = \mathbf{c}^\top W \mathbf{e} \approx \mathbf{c}'^\top H^\top W H \mathbf{e}'$

transferred LM-Steer!

Transferring to Another LM



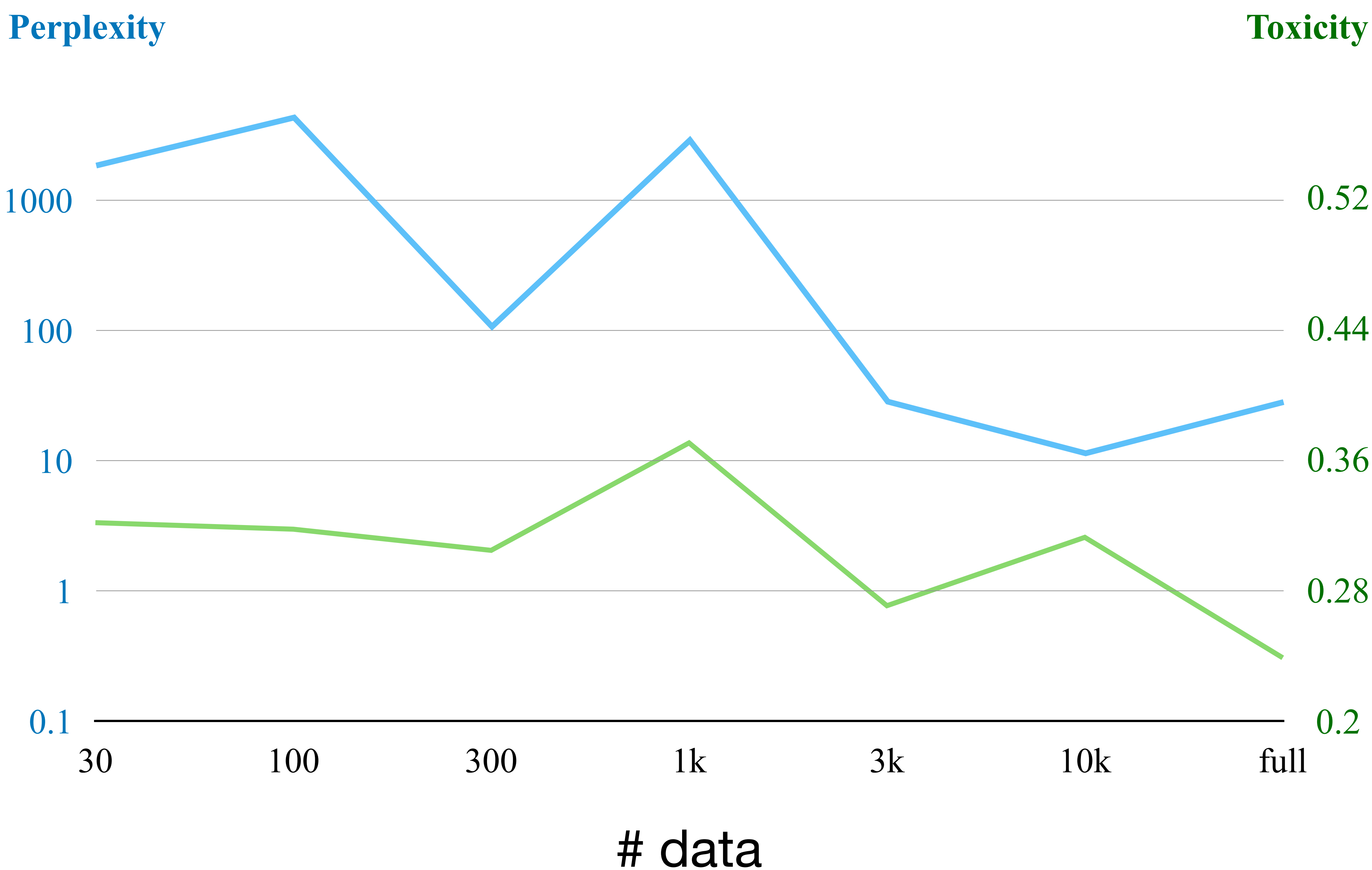
transfers about half of the detoxification capability

Computational Efficiency

	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**

- training only 0.9% of LM training parameters
- Marginal time overhead. Can be further reduced to 1.0 if the steering value ϵ is fixed.

Data Efficiency

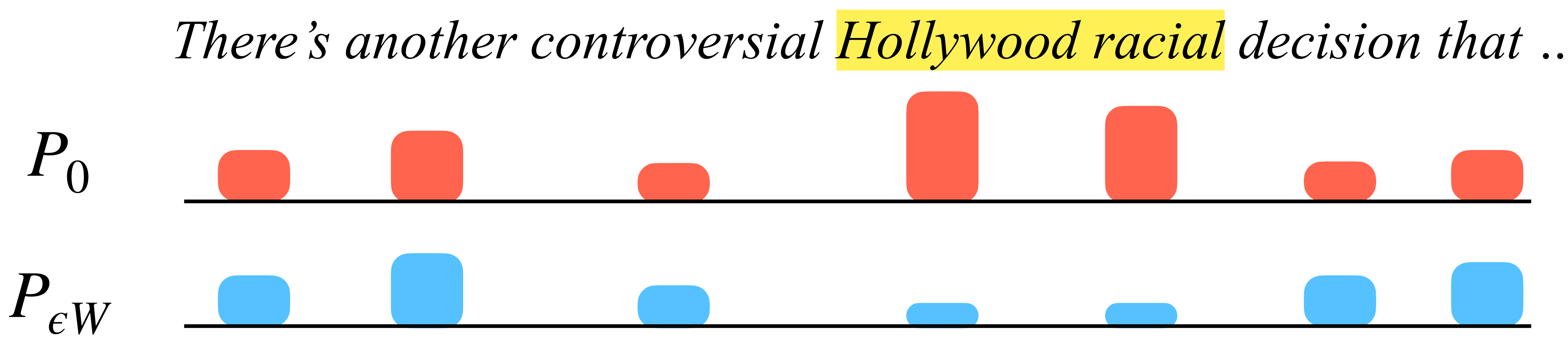


Highlighting Keywords

There's another controversial Hollywood racial decision that Stacey Dash is sinking her teeth into.
The UFC champ then suggested Justino is a longtime PED user with her most d**ning comments.
But I really have a question for you: Why would I go on a game show and play into the bulls**t allowing myself to be ranked by some fake competition?
I think sexism prevents this from being a real win for fat people.
If they want to be fair and non hypocritical idiots they should.

- Automatically highlighting text spans most related to a distribution.
- Example: toxic word highlighting by learning detoxification

Highlighting Keywords



- Motivation: what words are more likely in P_0 instead of P_W?
- Objective: looking for the text spans with the maximal sum of log-likelihood differences
- Inputs: sequences P_0 and P_W, #spans to look for n, max span length l
- Algorithm: dynamic programming

A Probe on the Word Embedding Space

SVD decomposition reveal words that are mostly related to a learned LM-Steer

SVD decomposition

$$\Delta logit(\mathbf{c}, \mathbf{e}) = \epsilon \mathbf{c}^\top W \mathbf{e} = \epsilon \mathbf{c}^\top U \Sigma V \mathbf{e}$$

$$= \epsilon \sum_i \sigma_i (\mathbf{c}^\top \mathbf{u}_i)(\mathbf{v}_i^\top \mathbf{e})$$

Each row $\mathbf{v}_i^\top$ in right matrix V looks for a dimension in the word embedding space, with decreasing significance σ_i

A Probe on the Word Embedding Space

Dim.	Matched Words
0 personal	mor, bigot, Stupid, retarded, coward, stupid, loser, clown, dumb, Dumb, losers, stupidity, garbage
1 abuses	stupid, idiot, Stupid, idiots, jerk, pathetic, suck, buff, stupidity, mor, damn, ignorant, fools, dumb
3	idiot, godd, damn, curses
5	Balk, lur, looms, hides, shadows, Whites, slippery, winds
7	bullshit, fiat, shit, lies, injust, manipulation critiques
8 political	disabled, inactive, whip, emo, partisan, spew, bombed, disconnected, gun, failing, Republicans

(Some dimensions were omitted as they match non-English words)

Future Work

- Comparing with input word embeddings: what is related and what is different?
- Are other contextual representations steerable? Any detailed analysis?
    - "Extracting Latent Steering Vectors from Pretrained Language Models" https://arxiv.org/pdf/2205.05124
- Going beyond linear transformation
- Calling for a better theoretical framework for LMs