Causal Estimation of Memorisation Profiles

Pietro Lesci,[🛡] Clara Meister,[ETH] Thomas Hofmann,[ETH] Andreas Vlachos,[🛡] Tiago Pimentel[ETH]
[🛡]University of Cambridge, [ETH]ETH Zürich
{pl487, av308}@cam.ac.uk
{clara.meister, thomas.hofmann, tiago.pimentel}@inf.ethz.ch

ACL 25 best paper

估计模型的Memorisation profile，模型在整个训练过程中记忆的趋势（每一条训练数据的）

(i) 在更大模型中更强且更持久
(ii) 受数据顺序和学习率影响
(iii) 在不同模型尺寸中具有稳定的趋势，因此可以从小模型预测大模型中的记忆情况

(i) 横坐标是checkpoint step (保存点)
(ii) 纵坐标是Treatment Step（处理数据批次）
(iii) 数据点，代表该checkpoint对该训练数据批次的记忆强度

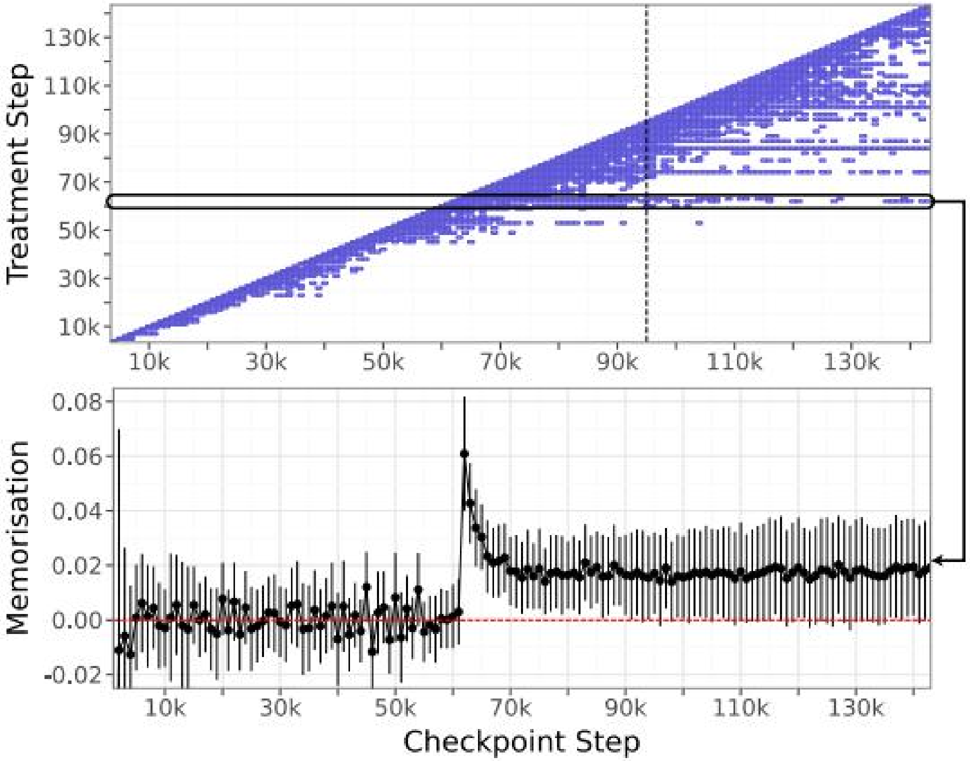


Figure 1: Memorisation profile (top) and path (bottom) of Pythia 6.9B. Each entry represents the expected counterfactual memorisation of instances trained on at a specific timestep ("Treatment Step") across model checkpoints ("Checkpoint Step"). The dashed vertical line indicates the end of the first epoch.

使用经济学学中的因果论，定义记忆：

1. 模型在训练过程中观察到某个实例对其正确预测该实例能力的因果影响
2. 反事实记忆：如果模型未经过训练该实例，该模型在该训练实例上的表现将会如何
   1. 因：模型是否训练该实例
   2. 果：该模型在该训练实例上的表现将会如何

instances *x* are sequences drawn from a target (unknown) distribution *p*(***x***)

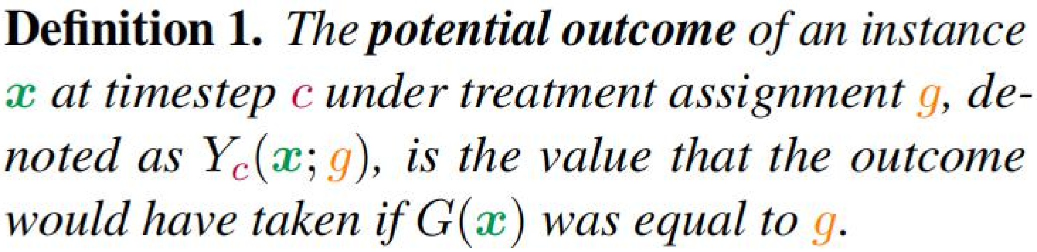
$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_{t-1}, \mathcal{B}_t)$$

c ∈ {0, 1, ..., T} 代表**checkpoint step**

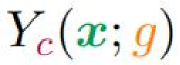
g ∈ {1, ..., T} ∪ {∞} 代表 **treatment step**，训练数据的步骤，∞代表没有用于训练，代表验证集

G(***x***) 代表 步骤*g* (当x被训练的步骤g)

$$Y_c(\boldsymbol{x}) \overset{\text{def}}{=} \gamma(\boldsymbol{\theta}_c, \boldsymbol{x}) = \log p_{\boldsymbol{\theta}_c}(\boldsymbol{x})$$ 量化具有参数 θc 的模型预测 x的能力

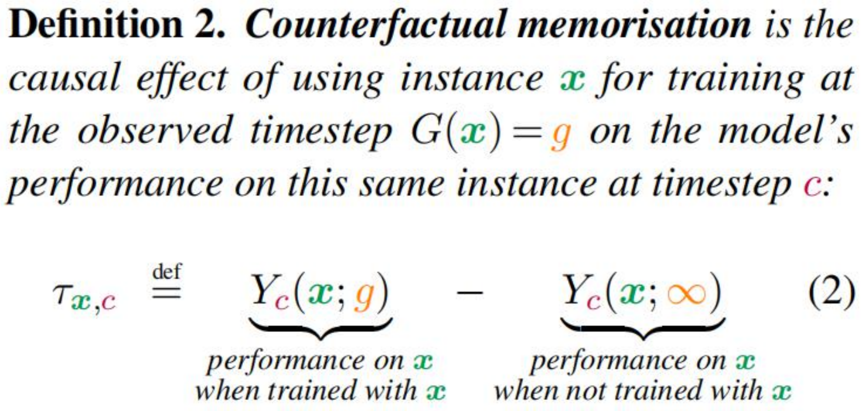


Definition 1. *The* ***potential outcome*** *of an instance $\boldsymbol{x}$ at timestep c under treatment assignment g, denoted as $Y_c(\boldsymbol{x}; g)$, is the value that the outcome would have taken if $G(\boldsymbol{x})$ was equal to g.*



$Y_c(\boldsymbol{x}; g)$

在checkpoint等于c时，模型处理实例x，在步骤g下的结果G(x)=g

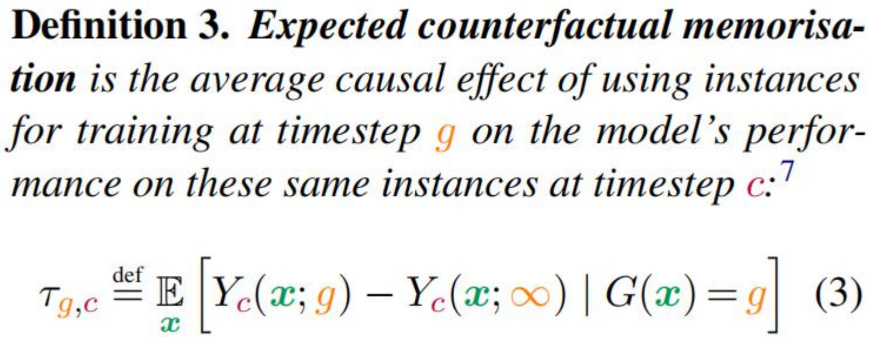


Definition 2. ***Counterfactual memorisation*** *is the causal effect of using instance $\boldsymbol{x}$ for training at the observed timestep $G(\boldsymbol{x}) = g$ on the model's performance on this same instance at timestep c:*

$$\tau_{\boldsymbol{x},c} \overset{\text{def}}{=} \underbrace{Y_c(\boldsymbol{x};g)}_{\substack{\textit{performance on } \boldsymbol{x} \\ \textit{when trained with } \boldsymbol{x}}} - \underbrace{Y_c(\boldsymbol{x};\infty)}_{\substack{\textit{performance on } \boldsymbol{x} \\ \textit{when not trained with } \boldsymbol{x}}} \tag{2}$$

用反事实记忆来估计模型及没记住

1. Y_c(x;g) 表示在训练时使用实例 x的表现，
2. Y_c(x; ∞) 表示在未使用实例 x进行训练时的表现。
3. 在相同checkpoint 下，用实例训过 – 没用实例训过的表现



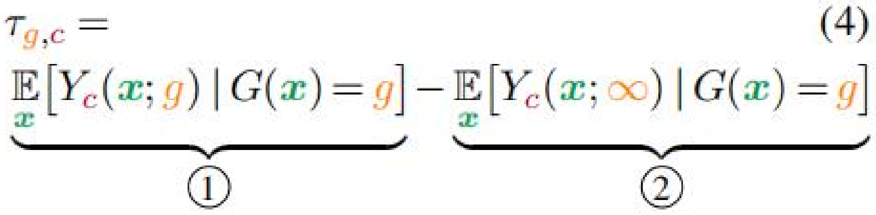
Definition 3. ***Expected counterfactual memorisation*** *is the average causal effect of using instances for training at timestep g on the model's performance on these same instances at timestep c:*[7]

$$\tau_{g,c} \overset{\text{def}}{=} \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};g) - Y_c(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = g\right] \tag{3}$$

实例表现的平均因果效应

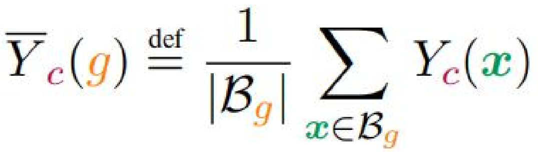
1. c < g: 不存在记忆现象
2. c == g: 瞬时记忆
3. c > g: 持久记忆
4. c = T: 残余记忆

反事实记忆定义如下:



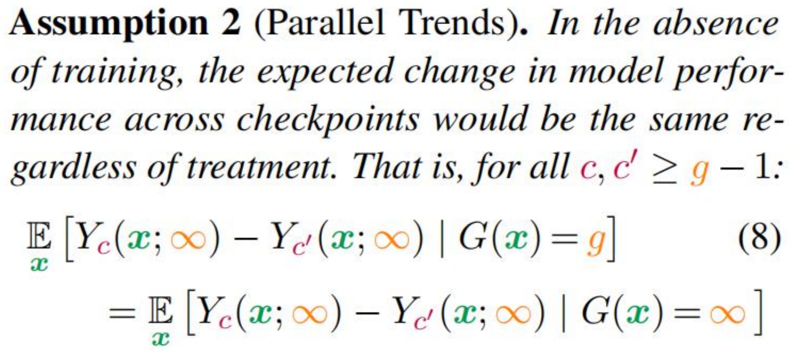
$$\tau_{g,c} = \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};g) \mid G(\boldsymbol{x})=g\right]}_{①} - \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};\infty) \mid G(\boldsymbol{x})=g\right]}_{②} \tag{4}$$

期望值 ① 可以从数据中直接估计



$$\overline{Y}_c(g) \overset{\text{def}}{=} \frac{1}{|\mathcal{B}_g|} \sum_{\boldsymbol{x}\in\mathcal{B}_g} Y_c(\boldsymbol{x})$$

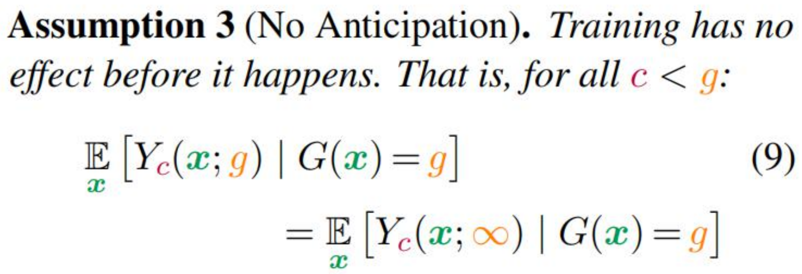
对于已经在时间步 g 被训练的实例 x，我们无法直接观测到如果模型没有在 g 训练该实例时的表现。期望②是反事实的，是无法直接估计的。



Assumption 2 (Parallel Trends)**.** *In the absence of training, the expected change in model performance across checkpoints would be the same regardless of treatment. That is, for all* $c, c' \geq g-1$*:*

$$\mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};\infty) - Y_{c'}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x})=g\right] \tag{8}$$
$$= \mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};\infty) - Y_{c'}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x})=\infty\right]$$

1. 如果模型没有见过实例 x，那么在任何两个check points之间，模型的性能变化趋势应该是相同的。
2. 假设如果模型**没有**在时间步 g 见过这个实例，那么它在 c和 c′之间的表现变化趋势应与那些**没有被见过**的实例的变化趋势是相似的。



Assumption 3 (No Anticipation)**.** *Training has no effect before it happens. That is, for all* $c < g$*:*

$$\mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};g) \mid G(\boldsymbol{x})=g\right] \tag{9}$$
$$= \mathbb{E}_{\boldsymbol{x}}\left[Y_c(\boldsymbol{x};\infty) \mid G(\boldsymbol{x})=g\right]$$

如果没训练该实例，训练在发生之前不会有任何影响

基于差分中的差分（DiD）设计的因果估计量，通过使用处理组与未处理组在结果随时间变化趋势上的差异来识别因果估计量

$$\begin{aligned}\tau^{\mathtt{did}}_{g,c} &= \mathop{\mathbb{E}}_{\boldsymbol{x}}[Y_c(\boldsymbol{x};g) - Y_{g-1}(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g] \quad (10)\\ &\quad - \mathop{\mathbb{E}}_{\boldsymbol{x}}[Y_c(\boldsymbol{x};\infty) - Y_{g-1}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = \infty]\end{aligned}$$

Estimator 2. *The **difference-in-differences estimator** (DiD), defined as:*

$$\widehat{\tau}^{\mathtt{did}}_{g,c} = \underbrace{\left(\overline{Y}_c(g) - \overline{Y}_{g-1}(g)\right)}_{\text{diff in trained}} - \underbrace{\left(\overline{Y}_c(\infty) - \overline{Y}_{g-1}(\infty)\right)}_{\text{diff in untrained}}$$

$$\begin{aligned}
&\tau^{\mathtt{did}}_{g,c}\\
&= \mathbb{E}\left[Y_c(\boldsymbol{x};g) - Y_{g-1}(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g\right] - \mathbb{E}\left[Y_c(\boldsymbol{x};\infty) - Y_{g-1}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = \infty\right] && \text{(21a)}\\
&= \mathbb{E}\left[Y_c(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g\right] - \underbrace{\mathbb{E}\left[Y_{g-1}(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g\right]}_{\text{no anticipation}} - \mathbb{E}\left[Y_c(\boldsymbol{x};\infty) - Y_{g-1}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = \infty\right]\\
& && \text{(21b)}\\
&= \mathbb{E}\left[Y_c(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g\right] - \underbrace{\mathbb{E}\left[Y_{g-1}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = g\right] - \mathbb{E}\left[Y_c(\boldsymbol{x};\infty) - Y_{g-1}(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = \infty\right]}_{\text{parallel trends}}\\
& && \text{(21c)}\\
&= \mathbb{E}\left[Y_c(\boldsymbol{x};g) \mid G(\boldsymbol{x}) = g\right] - \mathbb{E}\left[Y_c(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = g\right] && \text{(21d)}\\
&= \mathbb{E}\left[Y_c(\boldsymbol{x};g) - Y_c(\boldsymbol{x};\infty) \mid G(\boldsymbol{x}) = g\right] && \text{(21e)}\\
&= \tau_{g,c} && \text{(21f)}
\end{aligned}$$

模型数据：Pythia 模型套件，70M-12B，20TB数据，1.5个epochs，batch size = 1024, 一个epoch为95k

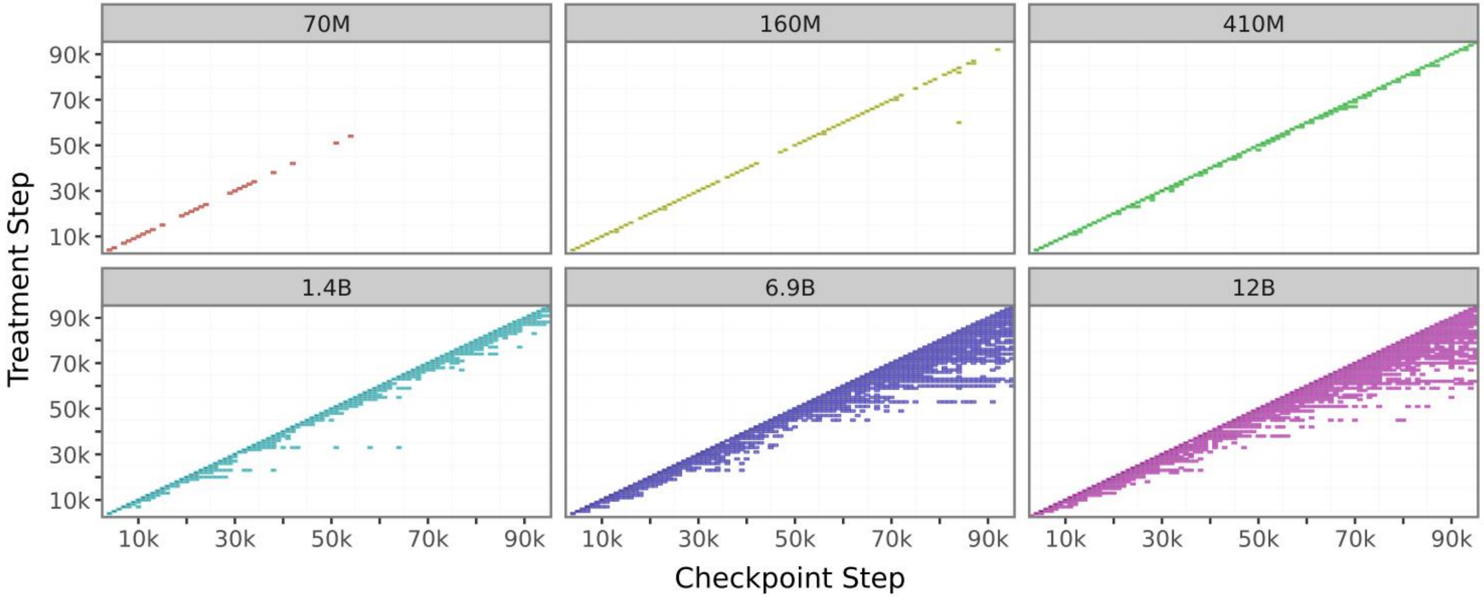


Figure 2: Memorisation profiles ($\tau_{g,c}$). We only show statistically significant entries.

(i) 在更大模型中更强且更持久
(ii) 即时记忆比训练后期更强
(iii) 在不同模型尺寸中具有稳定的趋势，因此可以从小模型预测大模型中的记忆情况

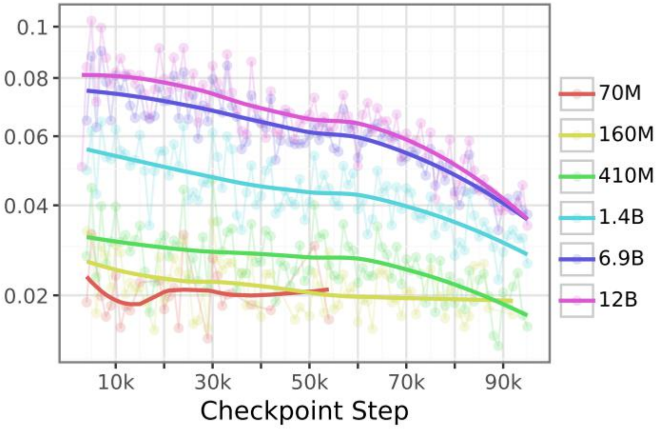


Figure 3: Instantaneous memorisation ($\tau_{g,c}$ for $g = c$). We only show statistically significant estimates.

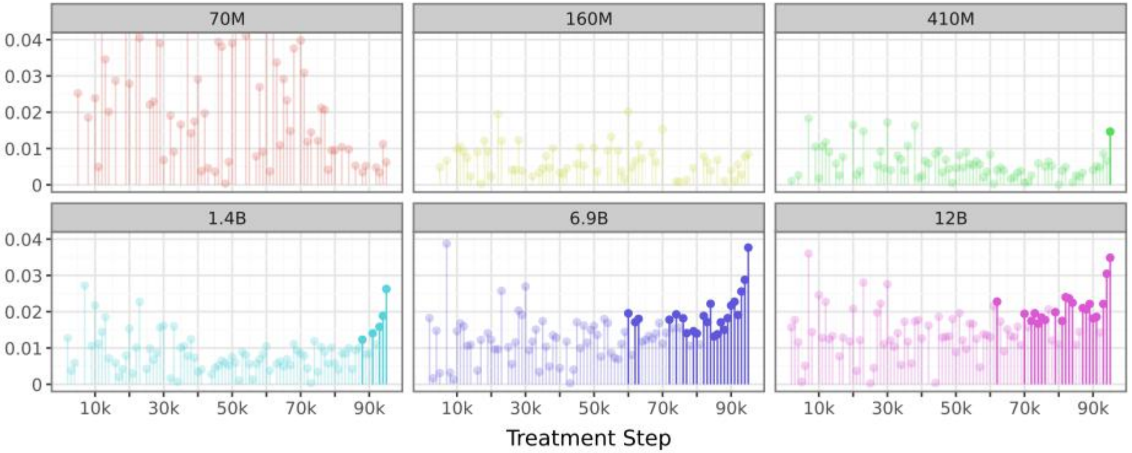
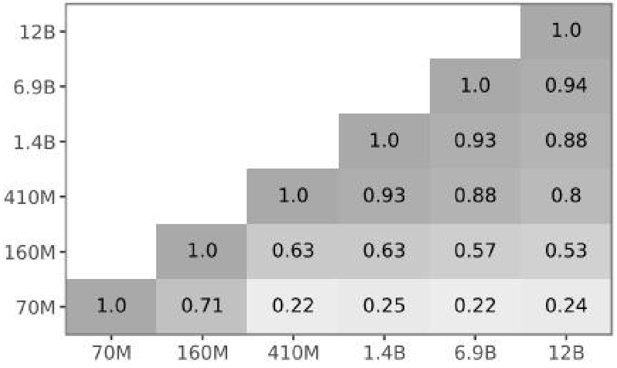


Figure 5: Residual memorisation ($\tau_{g,c}$ for $c = T = 95$k). Stronger colour intensity indicates statistical significance.

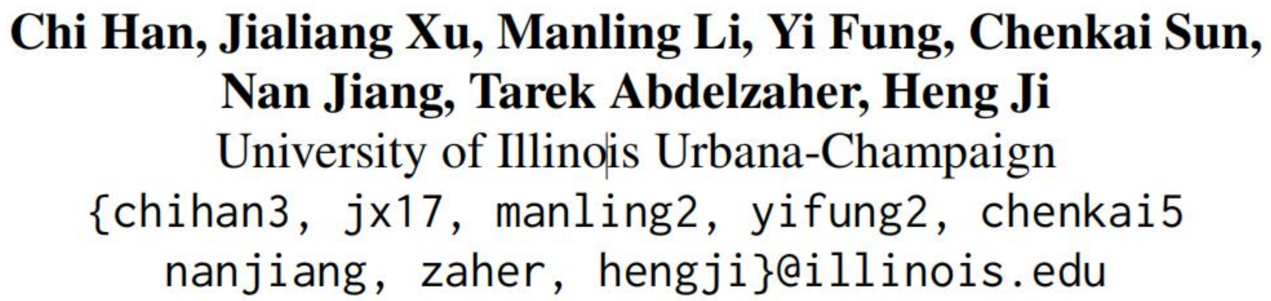
1. 即时记忆与余弦学习率调度相关
   1. 当学习率高时，优化过程将模型参数进一步推向局部最优方向，从而“覆盖”之前的信息并用新信息进行更新；这导致较高的即时记忆和较低的残余记忆
   2. 当学习率较低时，先前的信息被“遗忘”的较少，导致较高的残余记忆和较低的即时记忆
2. 较大的模型的记忆是可以通过较小的模型预测的



	70M	160M	410M	1.4B	6.9B	12B
12B						1.0
6.9B					1.0	0.94
1.4B				1.0	0.93	0.88
410M			1.0	0.93	0.88	0.8
160M		1.0	0.63	0.63	0.57	0.53
70M	1.0	0.71	0.22	0.25	0.22	0.24

Figure 6: Pearson correlation between the memorisation profile of different models.

Word Embeddings Are Steers for Language Models



Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun,
Nan Jiang, Tarek Abdelzaher, Heng Ji
University of Illinois Urbana-Champaign
{chihan3, jx17, manling2, yifung2, chenkai5
nanjiang, zaher, hengji}@illinois.edu

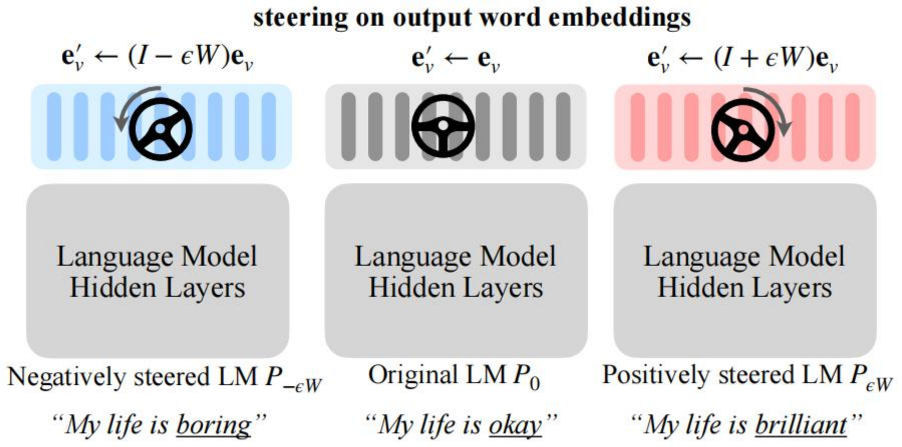
ACL 25 best paper

llm的预测token 的 logit 定义如下：

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

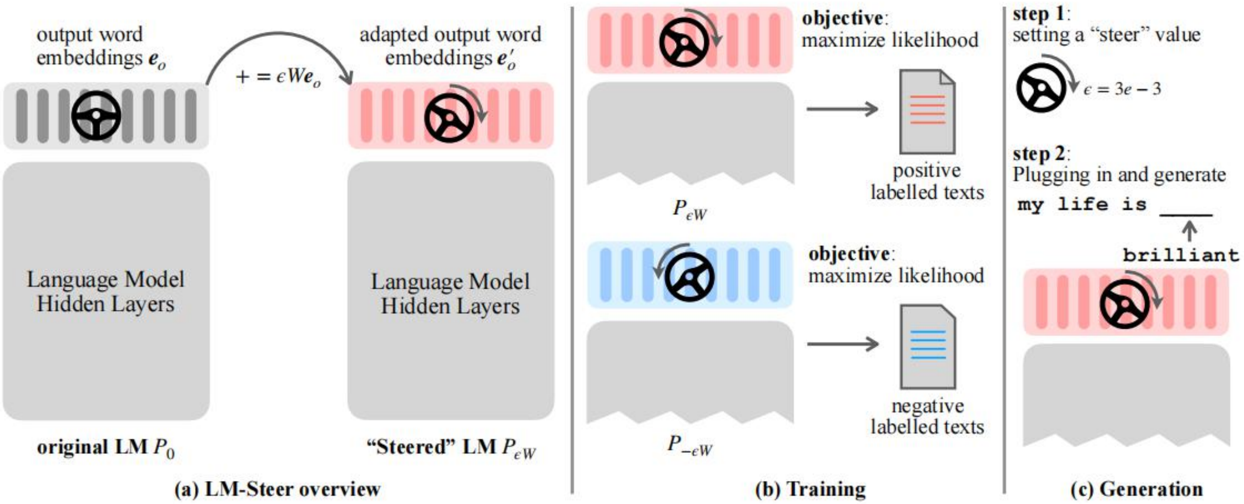
$$\mathbf{e}'_v = \mathbf{e}_v + \epsilon W \mathbf{e}_v = (I + \epsilon W)\mathbf{e}_v,$$

Theorem 1. *(Informal) With certain assumptions, shifting styles in language models is equivalent to a linear transformation in word embedding space.*

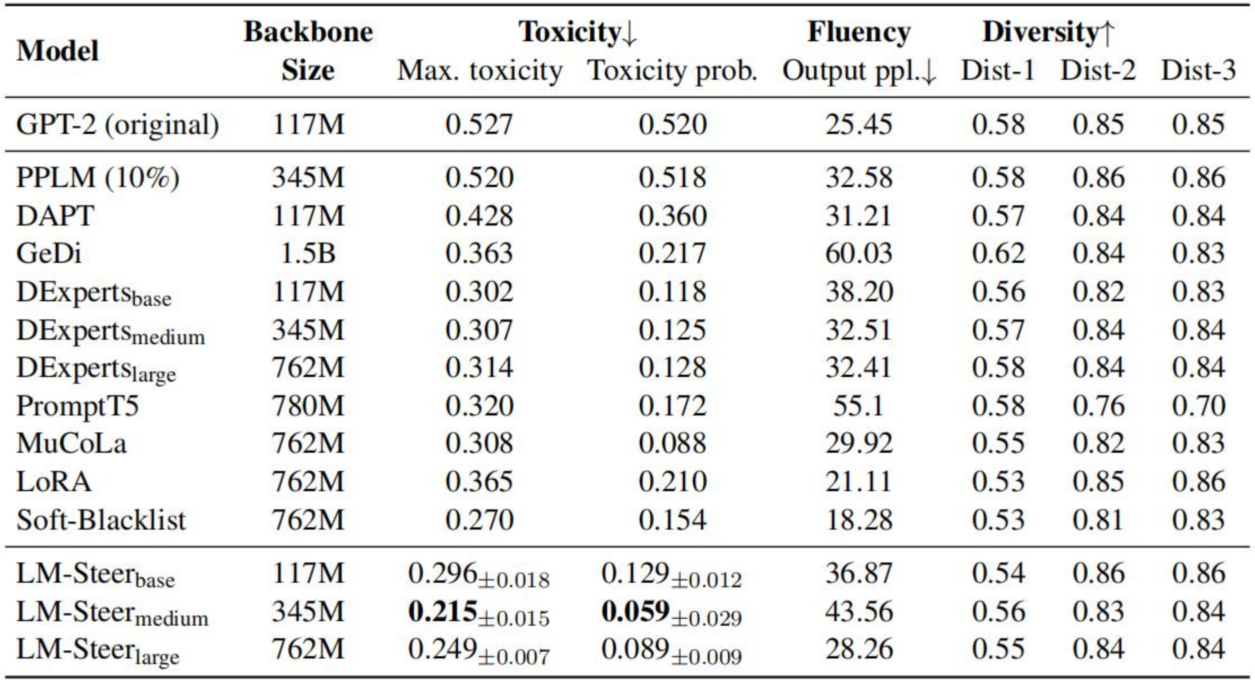


steering on output word embeddings
$\mathbf{e}'_v \leftarrow (I - \epsilon W)\mathbf{e}_v$
$\mathbf{e}'_v \leftarrow \mathbf{e}_v$
$\mathbf{e}'_v \leftarrow (I + \epsilon W)\mathbf{e}_v$
Language Model Hidden Layers
Language Model Hidden Layers
Language Model Hidden Layers
Negatively steered LM $P_{-\epsilon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

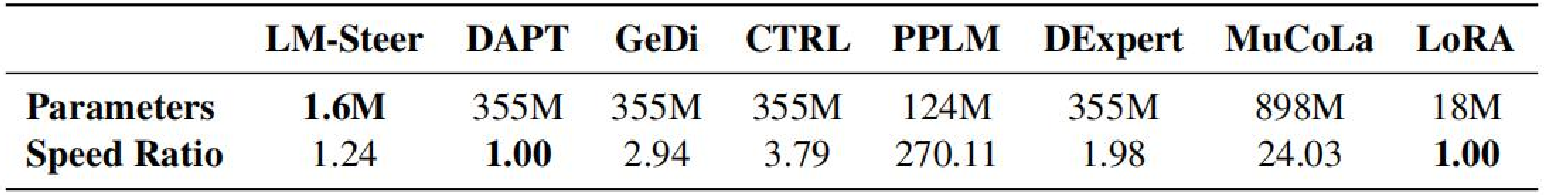
用output word embeddings操控模型语言风格
LM-Steers



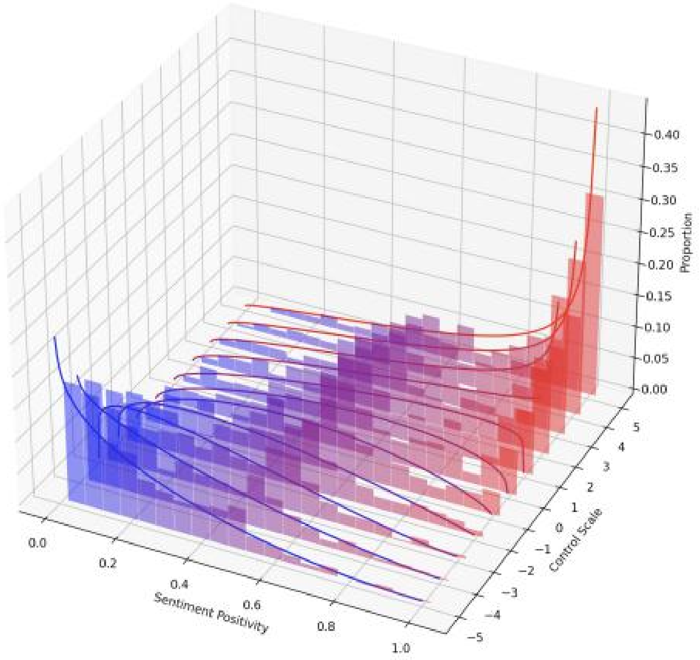
output word embeddings e_o
$+ = \epsilon W e_o$
adapted output word embeddings e'_o
Language Model Hidden Layers
Language Model Hidden Layers
original LM P_0
"Steered" LM $P_{\epsilon W}$
(a) LM-Steer overview
objective: maximize likelihood
$P_{\epsilon W}$
positive labelled texts
objective: maximize likelihood
$P_{-\epsilon W}$
negative labelled texts
(b) Training
step 1: setting a "steer" value
$\epsilon = 3e-3$
step 2: Plugging in and generate
my life is ____
brilliant
(c) Generation



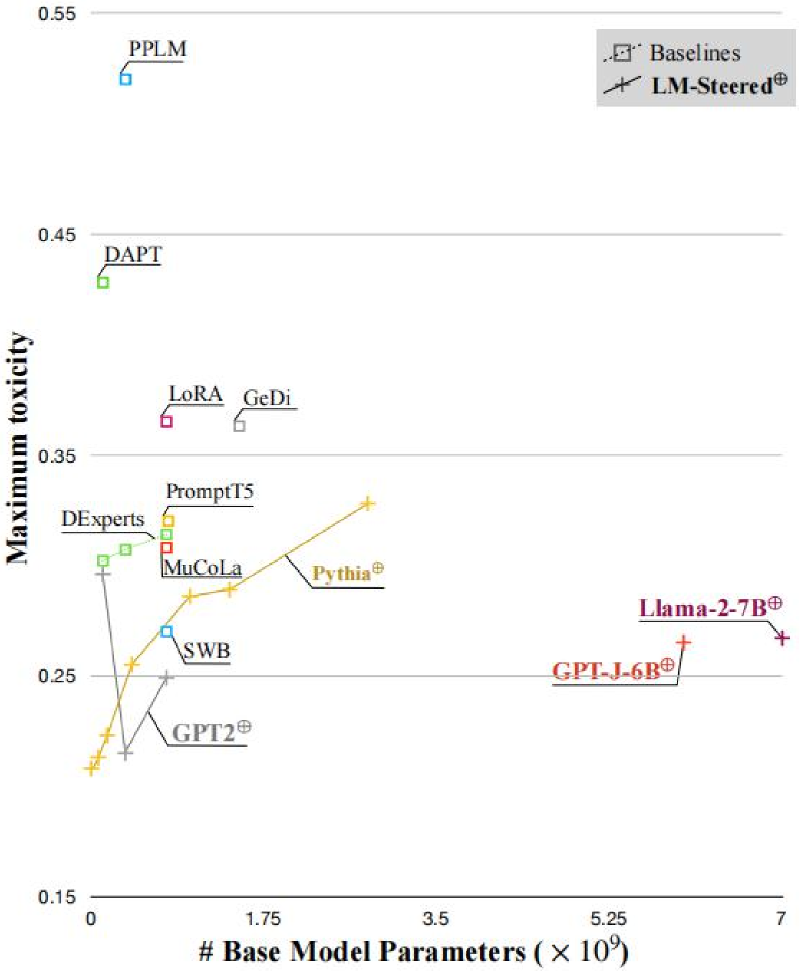
Model	**Backbone**	**Toxicity↓**		**Fluency**	**Diversity↑**		
	Size	Max. toxicity	Toxicity prob.	Output ppl.↓	Dist-1	Dist-2	Dist-3
GPT-2 (original)	117M	0.527	0.520	25.45	0.58	0.85	0.85
PPLM (10%)	345M	0.520	0.518	32.58	0.58	0.86	0.86
DAPT	117M	0.428	0.360	31.21	0.57	0.84	0.84
GeDi	1.5B	0.363	0.217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

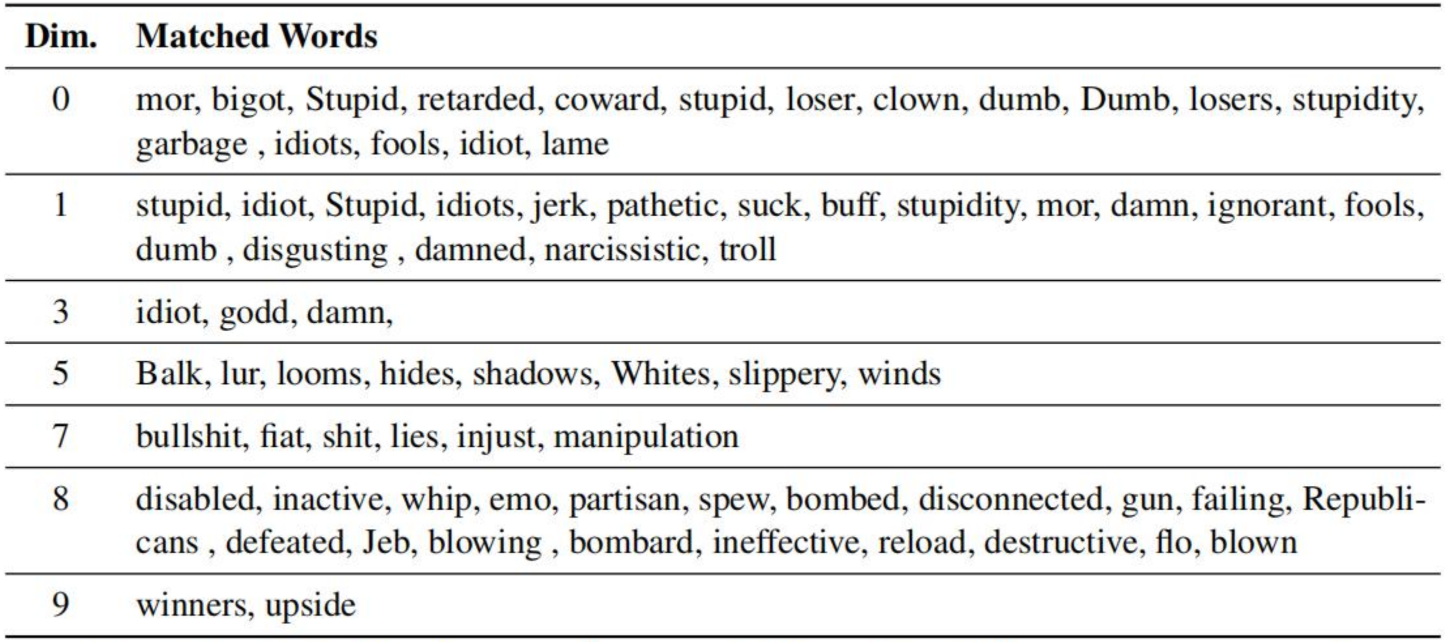


(a) Continuous control on sentiment with ϵ in $-5\epsilon_0 \sim 5\epsilon_0$ results in a sentiment distribution shift. Color indicates sentiment and height indicates frequency/density.



有一定的物理含义：

1. 风格转换解释embedding含义：*what word dimensions contribute to or contrast to a specific style*
   1. 对学习到的 W 进行了 SVD 分解，取 D(最重要的维度)转为token，哪些token在风格变换中的作用被放大



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

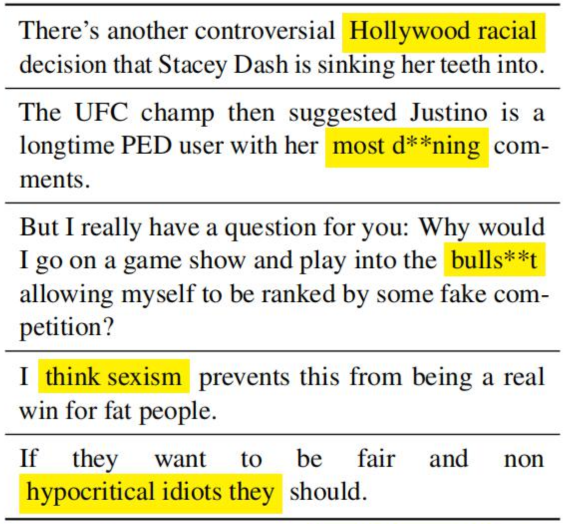
Table 6: Word embedding dimensions that are most influenced by LM-Steer on detoxification task.

有一定的物理含义：

2. 在风格化文本中突出显示关键词：*what specific words are most indicative of the style in a given sentence*

2.1 具体的文本中自动识别和突出显示那些最能指示文本风格的词语

$$\log P_{\epsilon W}(v_i|v_{i-1}\cdots v_{i-1}) - \log P_0(v_i|v_{i-1}\cdots v_{i-1})$$



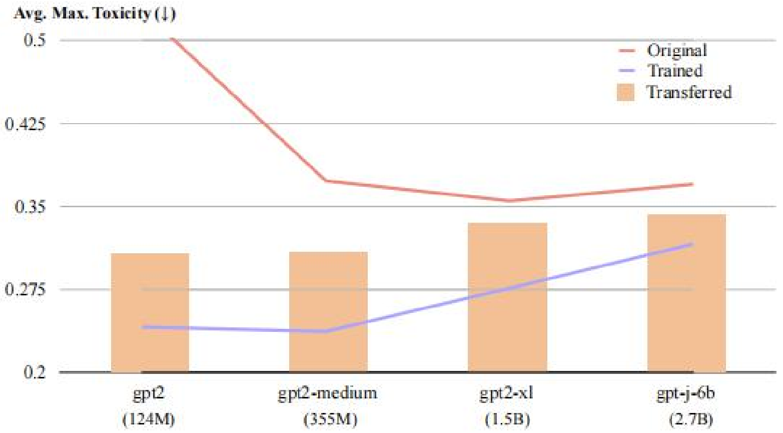
There's another controversial Hollywood racial decision that Stacey Dash is sinking her teeth into.
The UFC champ then suggested Justino is a longtime PED user with her most d**ning comments.
But I really have a question for you: Why would I go on a game show and play into the bulls**t allowing myself 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.

Table 7: Toxic sentences with toxic keywords highlighted by LM-Steer after training detoxification on GPT2-Large.

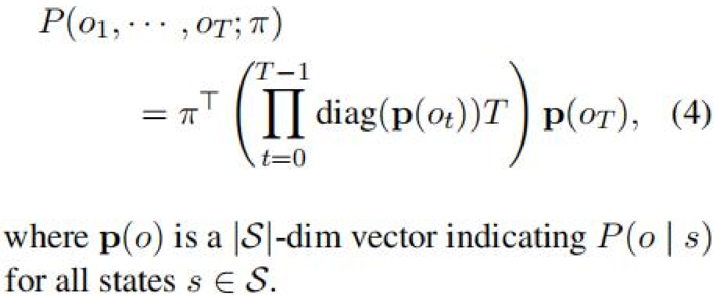
3. 在模型之间迁移 LM-Steer

$$\mathbf{e}_v = H\mathbf{e}'_v$$

$$\mathbf{c}^\top W\mathbf{e}_v = (H\mathbf{c}')^\top W(H\mathbf{e}'_v) = \mathbf{c}'^\top(H^\top WH)\mathbf{e}'_v$$



(a) Transferring an LM-Steer to other LMs with explicit-form calculation. The transferred LM-Steer maintains the detoxification ability partially.

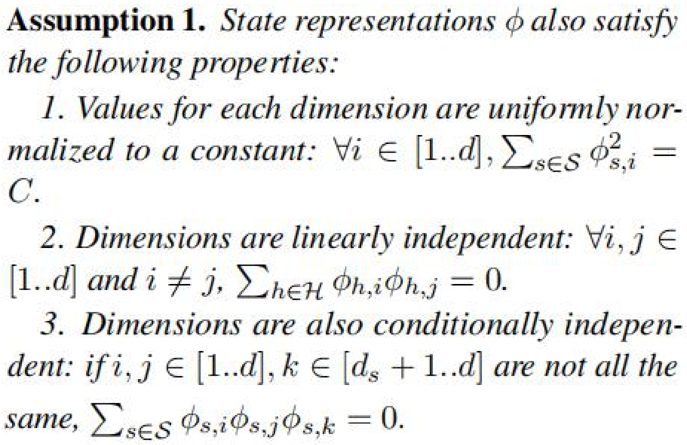


$$P(o_1, \cdots, o_T; \pi) = \pi^\top \left(\prod_{t=0}^{T-1} \mathrm{diag}(\mathbf{p}(o_t)) T \right) \mathbf{p}(o_T), \quad (4)$$

where $\mathbf{p}(o)$ is a $|\mathcal{S}|$-dim vector indicating $P(o \mid s)$ for all states $s \in \mathcal{S}$.

$$T(s, s') = \phi_{s,\text{semantic}}^\top A' \phi_{s',\text{semantic}} + \phi_{s,\text{condition}}^\top \phi_{s',\text{condition}}$$

$$\phi_s = \begin{pmatrix} \phi_{s,\text{semantic}} \\ \phi_{s,\text{condition}} \end{pmatrix}.$$

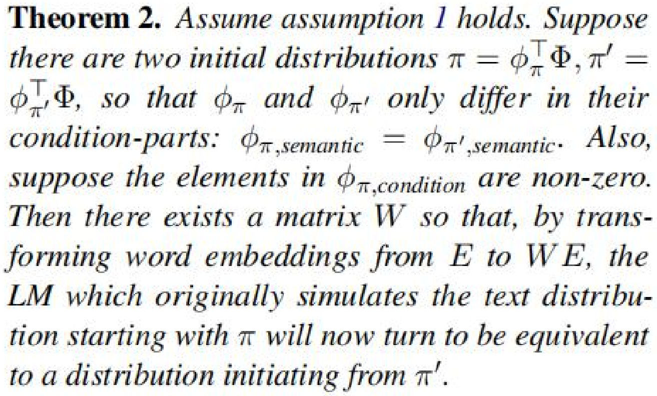


Assumption 1. *State representations ϕ also satisfy the following properties:*

1. Values for each dimension are uniformly normalized to a constant: $\forall i \in [1..d], \sum_{s\in\mathcal{S}} \phi_{s,i}^2 = C$.

2. Dimensions are linearly independent: $\forall i, j \in [1..d]$ and $i \neq j$, $\sum_{h\in\mathcal{H}} \phi_{h,i}\phi_{h,j} = 0$.

3. Dimensions are also conditionally independent: if $i, j \in [1..d], k \in [d_s + 1..d]$ are not all the same, $\sum_{s\in\mathcal{S}} \phi_{s,i}\phi_{s,j}\phi_{s,k} = 0$.



Theorem 2. *Assume assumption 1 holds. Suppose there are two initial distributions $\pi = \phi_\pi^\top \Phi, \pi' = \phi_{\pi'}^\top \Phi$, so that ϕ_π and $\phi_{\pi'}$ only differ in their condition-parts: $\phi_{\pi,semantic} = \phi_{\pi',semantic}$. Also, suppose the elements in $\phi_{\pi,condition}$ are non-zero. Then there exists a matrix W so that, by transforming word embeddings from E to WE, the LM which originally simulates the text distribution starting with π will now turn to be equivalent to a distribution initiating from π'.*