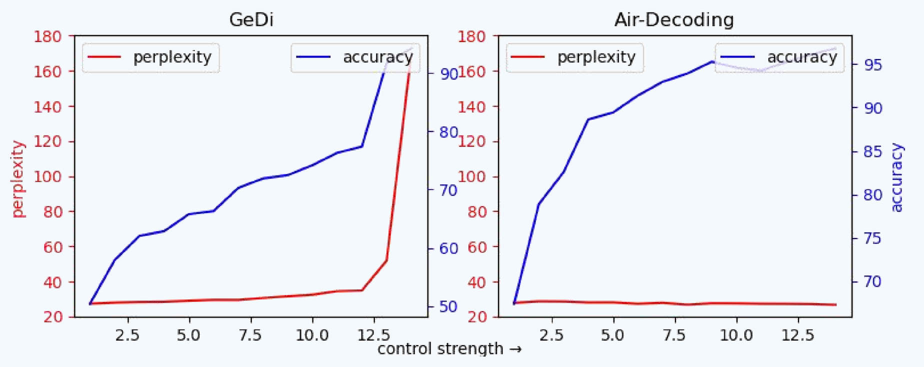
Air-Decoding: Attribute Distribution Reconstruction for Decoding-Time Controllable Text Generation

Tianqi Zhong[1], Quan Wang[2], Jingxuan Han[1], Yongdong Zhang[1], Zhendong Mao[1*]
1: University of Science and Technology of China;
2: MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications.

Motivations:

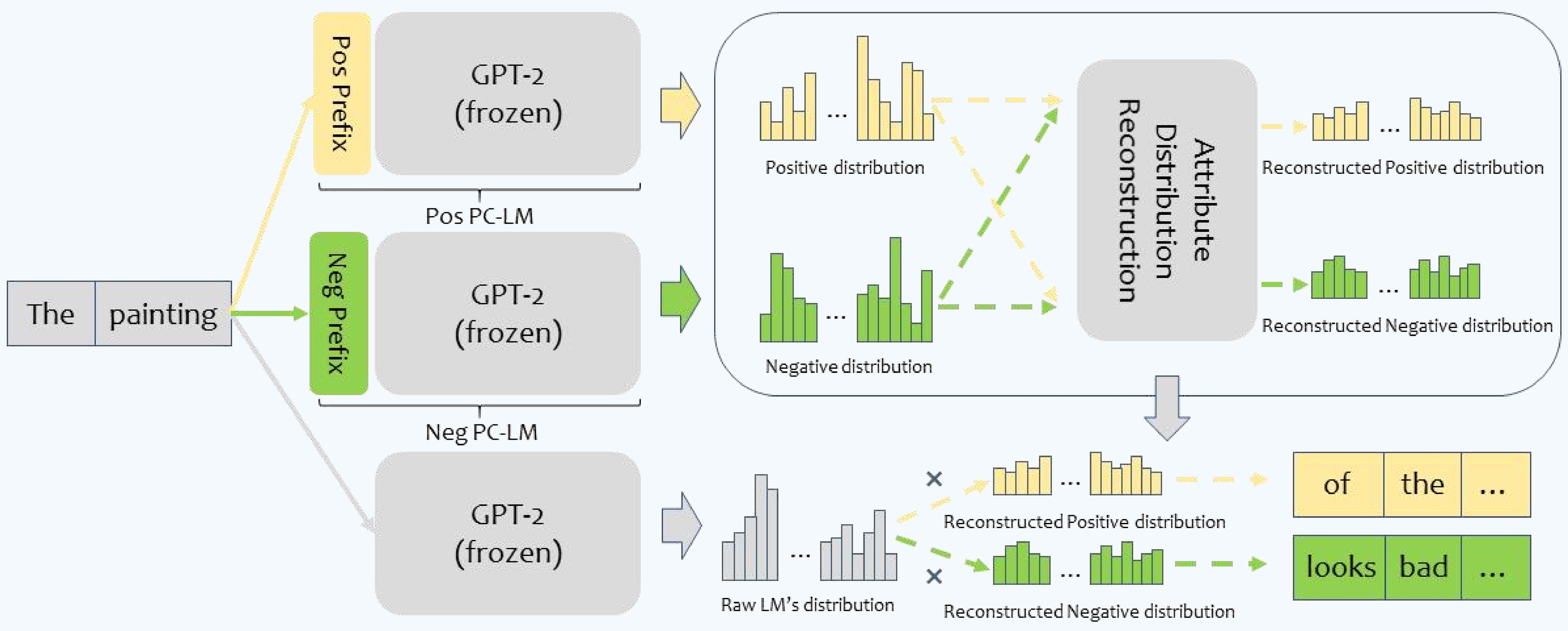


Prior work[1] has classified controllable text generation into three main categories as following:

1. **Retraining the whole parameters of CLMs**: Impressive control effects but large computational cost when the size of CLMs grows.
2. **Fine-tuning prefixes or prompts**: Low computational cost and fast inference speed but typically poor generalization.
3. **Decoding-Time approaches**: Good generalization and remarkable control effects but bad fluency under high control effects.

Attribute Collapse: The most severe issue with decoding-time methods is Attribute Collapse, which refers to the phenomenon that when the control strength increases to a certain critical value, the fluency of the generated text will rapidly decrease like the figure of GeDi[2] above. Therefore, how to solve the Attribute Collapse problem is a crucial issue.

Proposed Methods:



- **Preliminary**

Decoding-Time CTG can be formulated as following three equations, where $X_{1:T-1}$ is the given prompt, a is the desired attribute, and ω is the control strength that is an additive item.

- $P(x_{T:N}|x_{1:T-1},a)=\Pi_{t=T}^{N}P(x_t|x_{<t},a)$
- $P(x_t|x_{<t},a)\propto P(a|x_{1:t})^{\omega}P(x_t|x_{<t}),\ \ t\geq T$
- $P(a|x_{1:t})=\frac{P(a)\Pi_{j=T}^{t}P_{\phi_a}(x_j|x_{<j},a)}{\sum_{a'\in\{a,\bar{a}\}}\Pi_{j=T}^{t}P(a')P_{\phi_{a'}}(x_j|x_{<j},a')}$

- **Attribute Distribution via PC-LM**

We optimize two prefixes using dataset with corresponding attributes using language model loss as:

- $L_{LM}=-\sum_{k=1}^{K}log\,P_{\lambda,\theta_{a'}}(x_k|x_{<k},H_{\theta_{a'}})$

- **Attribute Distribution Reconstruction**

We design an attribute reconstruction method to make the distributions obtained by PC-LMs more balanced. First, we regularize the obtained attribute distribution before generating the next token x_t each time. Then we calculate $P(a|x_{1:t})$ using regularized $\widetilde{P}_{\lambda,\theta_{a'}}(*)$. The $\widetilde{P}_{\lambda,\theta_{a'}}(*)$ and the final decoding statement are formulated as:

- $\widetilde{P}_{\lambda,\theta_{a'}}\left(x_t\middle|x_{<t},H_{\theta_{a'}}\right)=-\frac{1}{\ln\left(P_{\lambda,\theta_{a'}}\left(x_t\middle|x_{<t},H_{\theta_{a'}}\right)\right)},\ \ a'\in\{a,\bar{a}\}$

- $P(x_t|x_{<t},a)=P(x_t|x_{<t})\left(\frac{\Pi_{j=T}^{t}\widetilde{P}_{\lambda,\theta_a}\left(x_j\middle|x_{<j},H_{\theta_a}\right)}{\sum_{a'\in\{a,\bar{a}\}}\Pi_{j=T}^{t}\widetilde{P}_{\lambda,\theta_{a'}}\left(x_j\middle|x_{<j},H_{\theta_{a'}}\right)}\right)^{\omega}$

Results From Main Experiments:

- **The main experimental results on IMDB dataset**

Method	Automatic Evaluation					Human Evaluation		
	Acc	PPL ↓	Dist-1	Dist-2	Dist-3	Rel.	Flu.	Top.
Pre-Tuning (Li and Liang, 2021)	62.15	38.49	0.11	0.53	0.82	2.28	3.52	2.81
Con Prefixes (Qian et al., 2022)	75.66	35.32	0.11	0.52	0.81	2.77	3.63	2.96
Discup* (Zhang and Song, 2022)	95.20	39.14	0.07	0.46	0.80	3.85	3.47	3.52
PPLM (Dathathri et al., 2019)	69.06	34.89	0.12	0.51	0.77	2.54	3.56	3.24
GeDi (Krause et al., 2021)	94.23	169.86	0.15	0.53	0.74	3.38	2.60	3.47
DExpert (Liu et al., 2021)	94.74	51.99	**0.16**	**0.65**	**0.85**	3.51	3.02	3.46
Air-Decoding (medium)	**96.82**	**26.66**	0.13	0.55	0.78	**4.03**	3.96	**3.85**
Air-Decoding (large)*	96.16	**18.59**	0.13	0.52	0.76	3.93	**4.01**	3.73

- **The main experimental results on AGNews dataset**

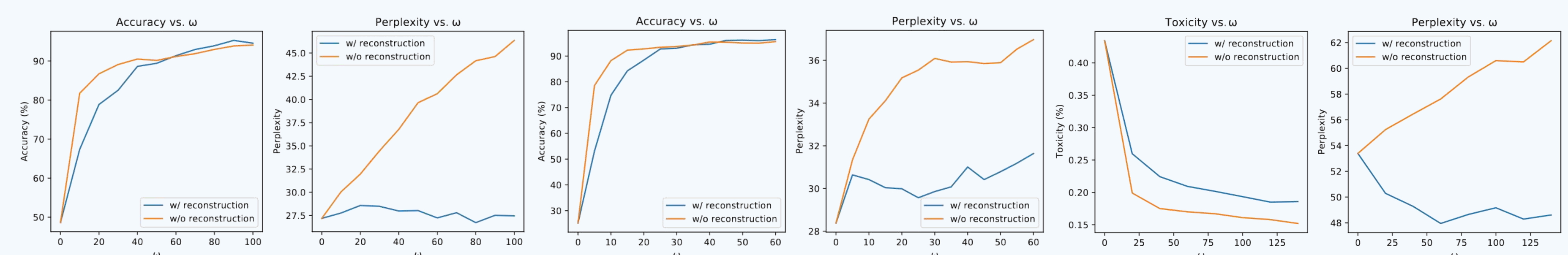
Method	Automatic Evaluation					Human Evaluation		
	Acc	PPL ↓	Dist-1	Dist-2	Dist-3	Rel.	Flu.	Top.
Pre-Tuning (Li and Liang, 2021)	72.74	64.43	0.09	0.49	0.74	2.85	3.05	2.84
Con Prefixes (Qian et al., 2022)	88.47	70.34	0.09	**0.50**	**0.75**	3.31	2.94	2.95
GeDi (Krause et al., 2021)	94.27	104.46	**0.10**	0.48	0.69	3.83	2.42	3.31
Air-Decoding (medium)	**97.21**	**31.18**	0.08	0.47	0.74	**4.07**	3.87	**3.80**
Air-Decoding (large)*	94.30	**22.31**	0.08	0.46	0.72	3.93	**3.94**	3.75

- **The main experimental results on Jigsaw dataset**

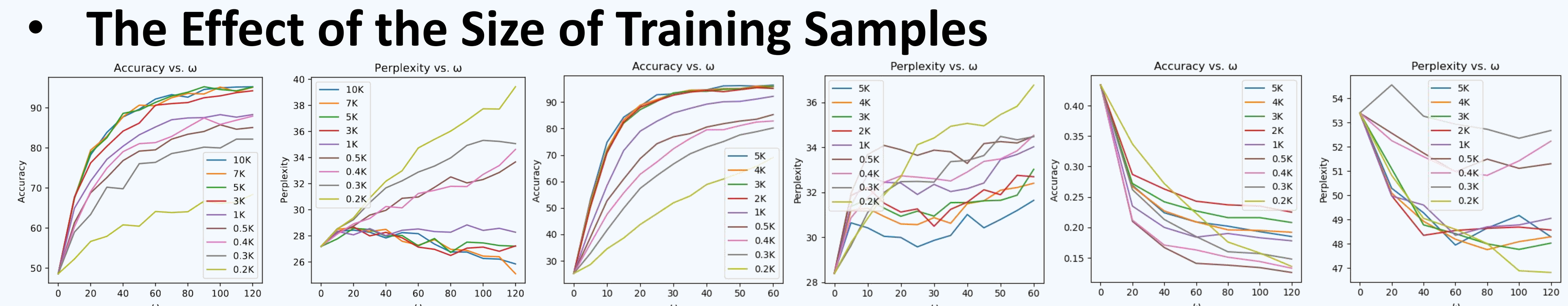
Method	Automatic Evaluation					Human Evaluation		
	Tox. ↓	PPL ↓	Dist-1	Dist-2	Dist-3	Rel.	Flu.	Top.
Pre-Tuning (Li and Liang, 2021)	49.2	92.20	0.07	0.40	0.68	2.24	2.37	2.93
Con Prefixes (Qian et al., 2022)	21.7	85.34	-	-	-	-	-	-
Discup* (Zhang and Song, 2022)	**14.8**	63.90	0.07	0.48	0.82	**3.90**	3.04	3.36
PPLM (Dathathri et al., 2019)	30.0	148.50	-	-	-	-	-	-
GeDi (Krause et al., 2021)	20.5	166.01	-	-	-	-	-	-
DExpert (Liu et al., 2021)	20.0	58.06	0.08	0.48	0.78	3.53	3.36	3.45
Air-Decoding (medium)	18.5	**48.29**	0.07	0.44	0.74	3.85	3.56	**3.74**
Air-Decoding (large)*	21.6	**38.86**	0.07	0.42	0.73	3.76	**3.64**	3.68

Further Analysis and Discussion:

- **The Effect of Distribution Reconstruction**



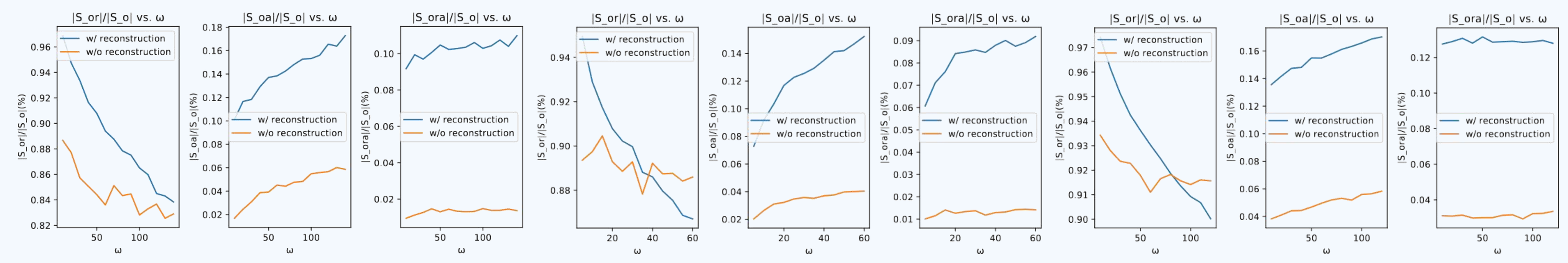
- **The Effect of the Size of Training Samples**



- **Analysis on Similarity between Distributions**

We define sets S_o, S_r, S_a as follows and define $S_{or}=S_o\cap S_r$, $S_{oa}=S_o\cap S_a$, $S_{ora}=S_o\cap S_r\cap S_a$. We use $|S|$ to denote the size of set S. Then we consider metrics: $\frac{|S_{or}|}{|S_o|}$, $\frac{|S_{oa}|}{|S_o|}$, $\frac{|S_{ora}|}{|S_o|}$.

- $P(x_t|x_{<t},a)$ denoted as d_o, $S_o\triangleq\{p_i|p_i\in d_o\wedge p_i\in TopK(d_o)\}$
- $P(x_t|x_{<t})$ denoted as d_r, $S_r\triangleq\{p_i|p_i\in d_r\wedge p_i\in TopK(d_r)\}$
- $P(a|x_{0:t})^{\omega}$ denoted as d_a, $S_a\triangleq\{p_i|p_i\in d_a\wedge p_i\in TopK(d_a)\}$



Reference:

[1]: A survey of controllable text generation using transformer-based pre-trained language models, Zhang, Hanqing, et al. ACM Computing Surveys'2023 https://arxiv.org/pdf/2201.05337
[2]: GeDi: Generative Discriminator Guided Sequence Generation, Krause, Ben, et al. EMNLP'2021 https://arxiv.org/pdf/2009.06367.pdf

Code and data available:

Code and Data: https://github.com/R1047/Air-Decoding

EMNLP 2023