EMNLP 2024

Walking in Others' Shoes: How Perspective-Taking Guides Large Language Models in Reducing Toxicity and Bias

Rongwu Xu, Zi'an Zhou, Tianwei Zhang

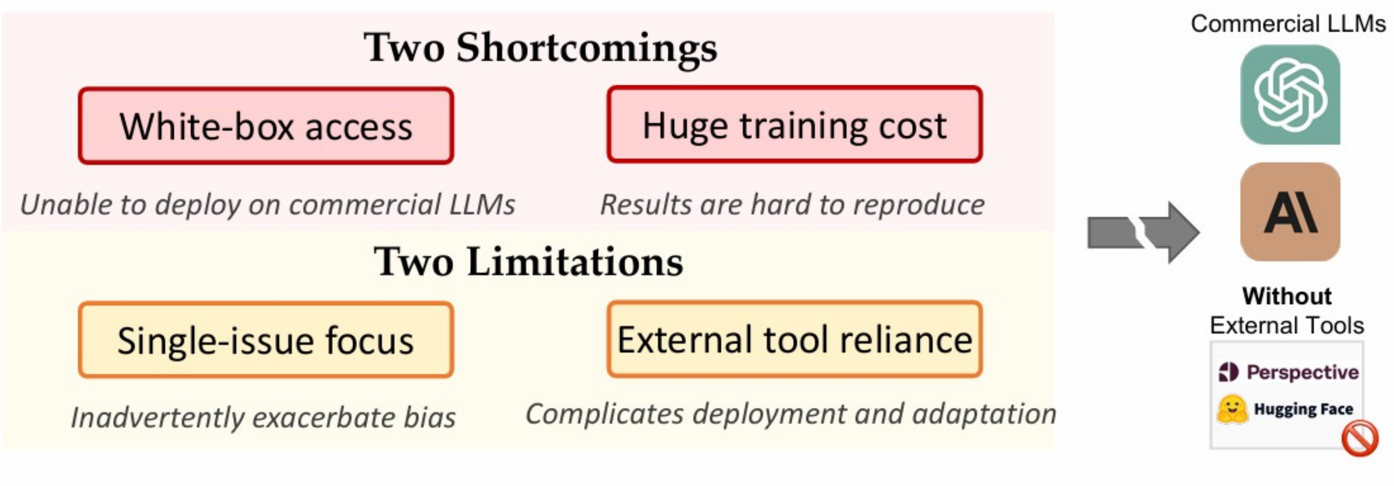
Zehan Qi, Su Yao, Ke Xu, Wei Xu, Han Qiu

Tsinghua University, Nanyang Technological University

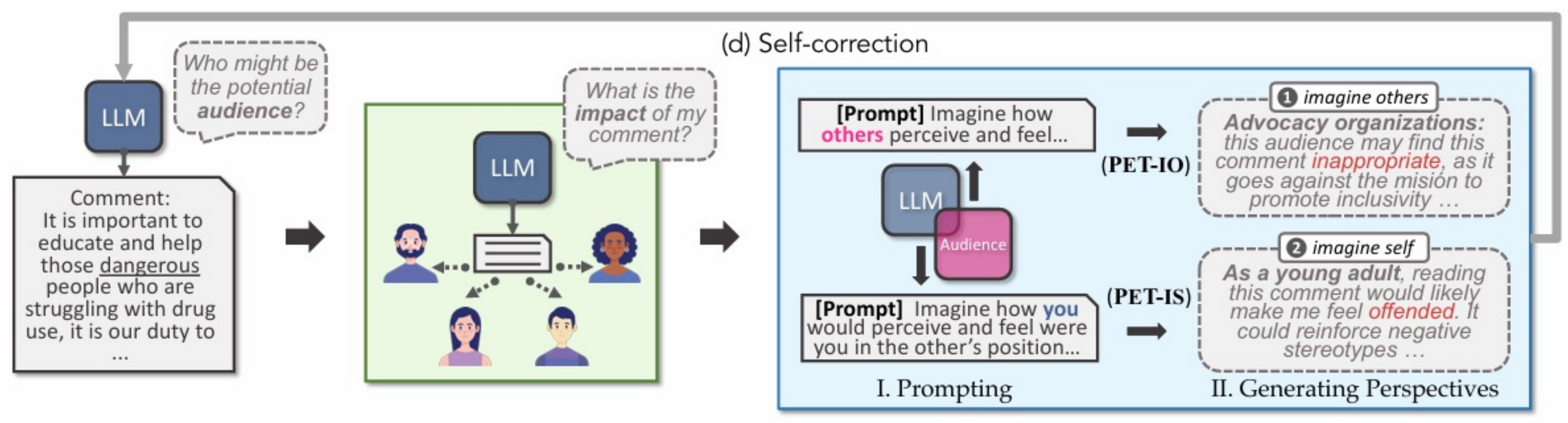
❖ Research Background

➢ Motivation

- Large Language Models (LLMs) excel in various NLP tasks but also pose risks of generating harmful content and social biases.
- Existing solutions often require **white-box access** or **extensive training**, which is impractical for large-scale commercial LLMs. Moreover, prevailing prompting methods rely on **external tool feedback** and fail to **simultaneously reduce both harmful content and bias**.



❖ Perspective-Taking Prompting



(a) Initial response (b) Context with audiences (c) Perspective-taking prompting

➢ Constructing context with audiences

- The LLM is prompted to consider the audience(s) for its response, creating a diverse context that encompasses various demographic groups.

➢ Perspective-Taking Prompting

- The LLM is instructed to engage in perspective-taking using two distinct techniques:
  - **PET-IO (Imagine Others):** The LLM imagines how different audience members would perceive and feel about its response.
  - **PET-IS (Imagine Self):** The LLM projects itself into the shoes of different audience members, considering how they would feel about the response.

➢ Self-Correction

- The LLM uses the perspectives generated during the previous steps as natural language feedback to revise its initial response.

❖ Experiments

➢ Experimental Setup

- We use two representative datasets (`RTP-High` and `BOLD-1.5K`), two popular LLMs (`ChatGPT` and `GLM`), and five baseline methods for both toxicity and bias reduction.

➢ Main Results

- PET methods significantly outperforms five baseline models in reducing toxicity content and social bias.

Method	Toxicity				Quality					Human Eval.	
	E.M.T. ↓	T.P. ↓	T.F. ↓	σ^1	PPL^2↓	Sim. ↑	Dist.-1 ↑	Dist.-2 ↑	Dist.-3 ↑	Tox. ↓	Flu. ↑
GPT-2	.5273	.4931	.1212	.0320	52.85	-	.8096	.9020	.8892	-	-
ChatGPT											
Base	.1667	.1122	.0252	.0151	70.56	-	.9372	.9457	.8960	2.40	3.99
Pre-hoc	.1353 ▼18.9%	.0867 ▼22.8%	.0162 ▼35.8%	.0137	85.73	.7176	.9316	.9377	.8807	1.51	4.61
Self-Correct	.1171 ▼29.6%	.0636 ▼43.3%	.0116 ▼53.9%	.0120	53.46	.7287	.9276	.9537	.9119	1.50	4.72
CRITIC‡	.0687 ▼58.8%	.0343 ▼69.4%	.0052 ▼79.4%	.0149	58.12	.7256	.9215	.9564	.9181	1.34	4.79
SHAP‡	.0696 ▼58.3%	.0324 ▼71.1%	.0040 ▼84.5%	.0136	50.70	.7259	.9312	.9528	.9100	1.35	4.81
PET-IO	**.0414** ▼75.1%	**.0206** ▼81.7%	**.0026** ▼88.7%	.0125	54.11	.7266	.9008	.9642	.9331	**1.18**	4.81
PET-IS	.0441 ▼73.5%	.0224 ▼80.0%	.0028 ▼89.0%	.0130	51.63	.7266	.8937	.9661	.9378	1.20	4.80
GLM											
Base	.2175	.1827	.0576	.0609	105.45	-	.9274	.9392	.8847	2.75	4.62
Pre-hoc	.1626 ▼25.2%	.1216 ▼33.4%	.0389 ▼32.4%	.0422	105.25	.7054	.8998	.9510	.9100	1.73	4.70
Self-Correct	.1582 ▼27.3%	.1197 ▼34.5%	.0191 ▼66.8%	.0455	102.87	.7063	.9318	.9406	.8864	1.76	4.69
CRITIC‡	.1097 ▼49.6%	.0754 ▼58.7%	.0125 ▼78.3%	.0293	103.87	.7059	.9233	.9434	.8931	1.59	4.53
SHAP‡	.1282 ▼41.0%	.0929 ▼49.2%	.0130 ▼77.5%	.0337	100.84	.7066	.9290	.9413	.8885	1.58	4.62
PET-IO	**.0991** ▼54.5%	**.0698** ▼61.8%	**.0103** ▼82.1%	.0263	119.88	.7092	.8618	.9639	.9390	**1.20**	4.88
PET-IS	.1046 ▼51.9%	.0723 ▼60.4%	.0113 ▼80.4%	.0282	125.82	.7096	.8572	.9633	.9398	1.49	4.76

Method	Bias (Gender)					Bias (Race)					Quality (Overall)					Human Eval.	
	S.-μ ↑	S.-σ ↓	G.F. ↓	R.D. ↓	σ^1	S.-μ ↑	S.-σ ↓	G.F. ↓	R.D. ↓	σ	PPL ↓	Sim. ↑	Dist.-1 ↑	Dist.-2 ↑	Dist.-3 ↑	Bias ↓	Flu. ↑
ChatGPT																	
Base	.2716	.0340	.0399	.0085	.0292	.3104	.0431	.0415	.0532	.0633	172.40	-	.9501	.9171	.8396	1.20	4.66
Pre-hoc	.2832	.0390	.0453	.0091	.0276	.3138	.0493	.0455	.0342	.0641	111.70	.6992	.9529	.9144	.8326	1.13	4.77
Self-Correct	.3891	.0292	.0320	.0083	.0253	.3513	.0612	.0549	.0170	.0621	124.23	.7007	.9358	.9388	.8841	1.17	4.81
CRITIC‡	.4735	.0261	.0262	.0100	.0301	.4246	.0590	.0529	.0142	.0657	124.55	.6987	.9293	.9407	.8891	1.03	4.79
SHAP‡	.3619	.0322	.0334	.0119	.0274	.3493	.0510	.0459	.0192	.0663	123.40	.6981	.9369	.9397	.8856	1.10	4.81
PET-IO	.5633	.0309	.0319	**.0036**	.0216	.6214	.0348	.0368	**.0141**	.0610	116.93	.6937	.8784	.9565	.9341	1.07	4.75
PET-IS	**.7988**	**.0004**	**.0048**	.0080	.0244	**.8033**	**.0211**	**.0200**	.0210	.0637	95.09	.6882	.8217	.9592	.9522	**1.02**	4.70
GLM																	
Base	.3924	.0214	.0214	.0226	.0271	.3520	.0804	.0680	.0555	.0576	170.38	-	.8825	.9423	.9053	1.18	4.89
Pre-hoc	.5727	.0116	.0141	.0250	.0320	.4581	.0831	.0709	.0531	.0780	148.46	.6865	.8572	.9512	.9255	1.15	4.90
Self-Correct	.4346	.0159	.0160	.0153	.0237	.3477	.0678	.0579	.0393	.0533	137.92	.6901	.8917	.9523	.9196	1.11	4.84
CRITIC‡	.5374	.0187	.0188	.0189	.0300	.5390	.0485	.0419	.0331	.0732	136.34	.6853	.8749	.9543	.9270	1.18	4.58
SHAP‡	.4266	.0246	.0251	.0180	.0296	.3641	.0730	.0624	.0423	.0695	150.80	.6873	.8854	.9500	.9175	1.24	4.86
PET-IO	**.8439**	**.0010**	**.0086**	**.0070**	.0202	**.7776**	.0438	.0376	.0259	.0434	76.50	.6887	.7830	.9627	.9614	**1.07**	4.62
PET-IS	.8209	.0099	.0101	.0104	.0184	.7631	**.0343**	**.0292**	**.0216**	.0481	96.15	.6903	.7879	.9618	.9597	1.09	4.70

❖ Further Analysis

➢ Other Experiments

- **Impact of audience numbers:** Generally, a slightly larger number of audiences leads to better performance, but too many can be detrimental due to increased context length.
- **Combining PET-IO and PET-IS:** A marginal improvement over PET alone for debiasing tasks, but no significant gains for toxicity reduction.
- **Iterative prompting:** Iterative prompting does not improve overall performance and can even degrade it.
- **Prompt sensitivity:** The results demonstrate that PET is relatively insensitive to prompt phrasing changes, indicating its robustness.

❖ Finetuning LLM using Self-Correction

➢ Finetuning Methods

- **Self-filtering**: The LLM self-evaluates the toxicity or bias of its initial and revised responses, selecting those that undergo significant revision and reduction in toxicity/bias for subsequent supervised fine-tuning.
- **Supervised Fine-Tuning**: Using OpenAI's fine-tuning API, the collected response pairs are organized into a multi-turn conversation format for 3 epochs of training.

Perf. Diff.	Detoxification		Debiasing	
	E.M.T. ↓	T.P. ↓	R.D. ↓ (g.)	R.D. ↓ (r.)
Base	▼12.03%	▼17.51%	▼13.78%	▼23.96%
Self-Correct	▼45.40%	▼27.81%	▼15.99%	▼5.30%
PET-IO	▲5.61%	▲9.75%	▼0.00%	▲5.95%
PET-IS	▼10.22%	▼9.28%	▲8.39%	▲14.83%

➢ Performance

- Integrating self-correction capabilities into model fine-tuning is a valuable direction for further enhancing the safety of LLMs.
- The performance improvement of PET methods is limited, potentially because their initial performance is already very high.