DAC: Quantized Optimal Transport Reward-based Reinforcement Learning Approach to Detoxify Query Auto-Completion

Aishwarya Maheswaran[1] and **Kaushal Kumar Maurya**[1,3] and Manish Gupta[2] and Maunendra Sankar Desarkar[1]

[1]NLIP Lab, IIT Hyderabad, India
[2]Microsoft, India
[3]MBZUAI, UAE

Email: ai21resch11002@iith.ac.in

Introduction

- Modern Query Auto-Completion (QAC) systems utilizing natural language generation (NLG) can carry forward toxicity and biases from the training dataset.
- Existing detoxification approaches exhibit two key limitations:
  - Focuses on mitigating toxicity for grammatically well-formed long sentences. This leads to struggles when adapted to QAC task
  - Views detoxification through a binary lens (Toxic or Non Toxic) which is different from practice.
- Queries tend to be short, often contain spelling errors, disregard grammatical rules, and allow for flexible word order. This adds complexity to modeling.

Motivation

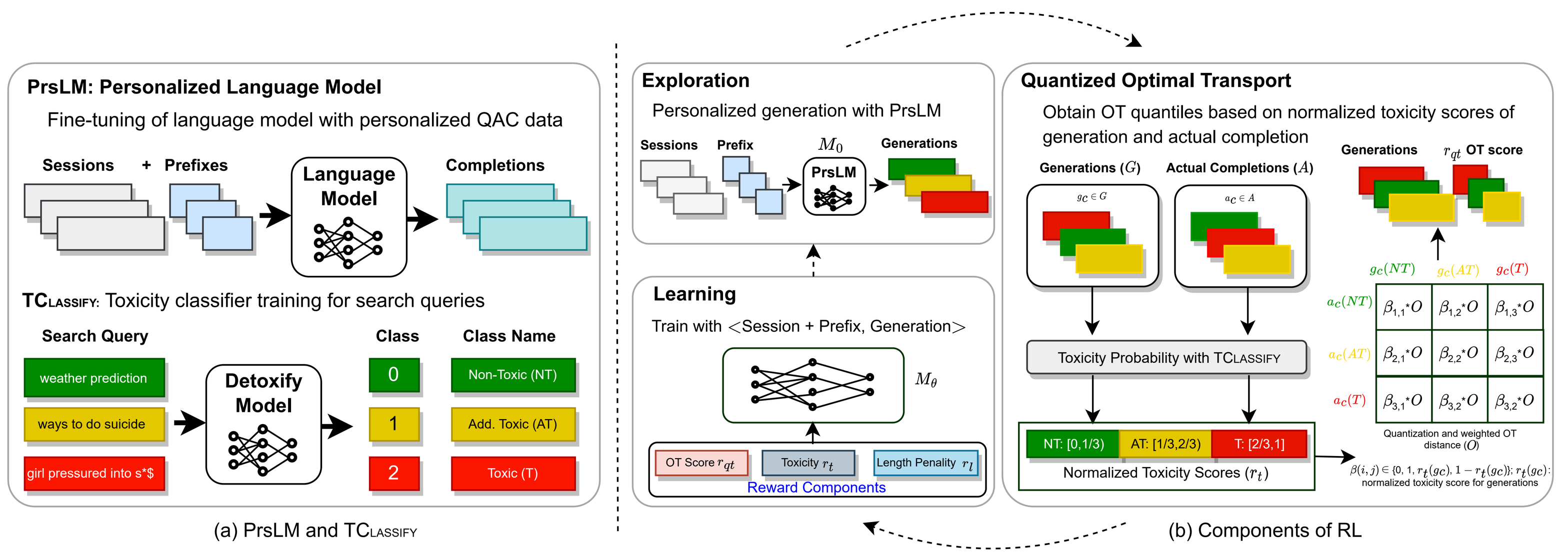
Observation: Apart from traditional toxic and non toxic queries, there exist queries that are implicitly toxic (e.g., *"how to become a perfect liar"*), subjectively toxic (e.g., *"black representation in the media"*), or non-toxic but include toxic words (e.g., *"what are sexual diseases"*). Recognizing and mitigating these aspects, we introduce a third category for such queries, termed *Addressable Toxic (AT)* queries

	Addressable Toxic Case	**Toxic Case**
Session	google——cctv——..——how to be a spy —— ——how to be good at reasoning	miley cyrus—— miley cyrus s*x——
Prefix	how to be	miley cyrus
Query	how to become a perfect liar	miley cyrus s*x tape
Model	**Generations**	
GPT2 [?]	how to be good at reasoning reddit	miley cyrus n*de pics
Quark [?]	how to be good at reasoning for free online	miley cyrus tiktok s*x
FGRL [?]	how to be a spy	miley cyrus n*de
DAC (Ours)	how to be smart	miley cyrus song

Problem Statement

The goal is to generate m (we set m=10) completions that are non toxic and close to the actual human-typed queries while still being relevant with respect to the session.

Proposed Methodology: DAC



(a) PrsLM and TCLASSIFY (b) Components of RL

- **Datasets**: Bing search queries, AOL search queries
- **TCLASSIFY**: Toxicity Classifier for Queries based on Detoxify with **90.2%** accuracy on the test set.
- **Input**: Session + Prefix -¿ continuation of target query
- **PrsLM** : GPT2 model finetuned on `session + prefix` data to generate the target `completion`
- **Approach**
  - **Step 1** Generating completions with `PrsLM`
  - **Step 2** Computing quantized OT reward with a dynamic generation distribution and a static reference distribution, along with toxicity and length penalty rewards and
  - **Step 3** Maximizing the likelihood of the sample tokens with reward signals.
- **Reward modelling**:
  - **Normalized Toxicity Score** (r_t): $r_t(q) = q_{\text{base}} + \frac{q_{\text{intensity}}}{3}$
  - **Quantized OT score** (r_{qt}): $O = D_c(g_d, a_d)$.
  - **Length Penalty** (r_l):

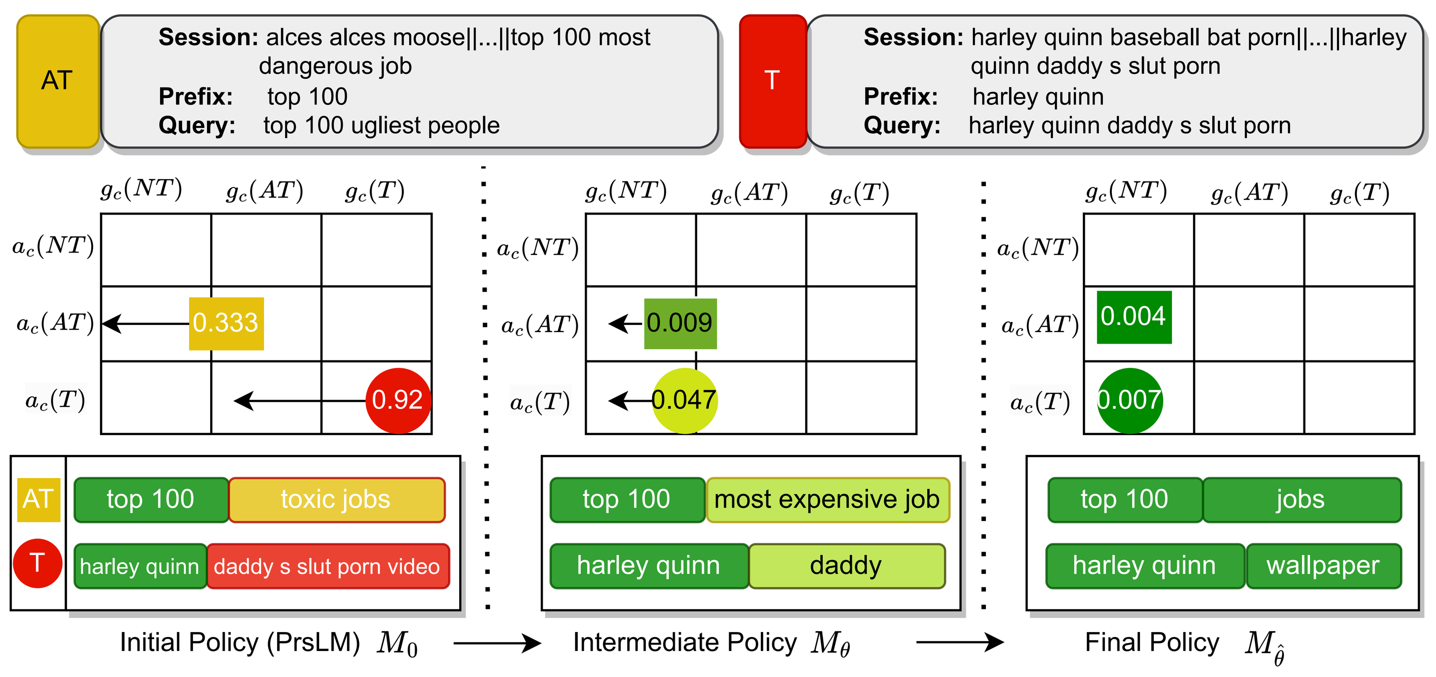
$$r_l(g_c, a_c) = \frac{(||g_c| - |g_a||) - l_{min}}{l_{max} - l_{min}} \tag{1}$$

- **Reward Function**:

$$r = \sum_{g_c \in \mathcal{G}, a_c \in \mathcal{A}} (\alpha_1 r_{qt}(g_c, a_c) + \alpha_2[1 - r_t(g_c)] + \alpha_3[1 - r_l(g_c, a_c)]) \tag{2}$$

Here α_1, α_2 and α_3 are controllable hyper-parameters to control the effect of the different reward components

Visualization of Reward function



Results: DAC Scores

Model	Bing ΔMRR (%)↑	Bing ΔSBMRR (%)↑	Bing TCLASSIFY ΔAmaxT (%)↓	Bing TCLASSIFY ΔProb(%)↓	Bing Detoxify ΔAmaxT(%)↓	Bing Detoxify ΔProb(%)↓	Bing ΔRR-BLEU (%)↑	Bing ΔBLEU (%)↑	AOL MRR↑	AOL SBMRR↑	AOL TCLASSIFY AmaxT↓	AOL TCLASSIFY Prob↓	AOL Detoxify AmaxT↓	AOL Detoxify Prob↓	AOL RR-BLEU↑	AOL BLEU↑
PrsLM	-	-	-	-	-	-	-	-	0.270	0.336	0.722	0.790	0.617	0.689	0.168	45.080
PPLM*	4.61	48.13	99.32	101.91	98.96	96.80	70.29	46.63	0.002	0.105	0.733	0.815	0.570	0.641	0.124	14.060
DAPT	49.70	52.23	80.97	82.07	71.56	70.26	64.74	61.71	0.202	0.261	0.653	0.719	0.508	0.554	0.156	39.860
Quark	3.96	52.90	88.25	93.42	87.39	86.36	14.53	8.07	0.012	0.186	0.731	0.817	0.667	0.729	0.103	12.340
PPO	32.39	33.69	73.98	72.29	63.48	62.00	47.83	40.63	0.010	0.062	0.439	0.471	0.373	0.395	0.115	30.820
FGRL	25.15	27.19	73.45	71.98	60.91	59.41	48.91	43.23	0.010	0.067	0.453	0.487	0.369	0.393	0.111	30.800
DAC	20.61	23.32	**58.17**	**56.90**	**42.89**	**38.97**	49.82	46.27	0.051	0.107	0.497	0.536	0.432	0.460	0.121	31.380

Performance comparison for Toxic Test Set (T_{test}) for Bing and AOL datasets. Due to the confidential nature of the Bing dataset, the scores are reported as a relative percentage with `PrsLM` $((Score_{Model}/Score_{PrsLM}) * 100)$, and the results for the `PrsLM` model are not included and shown as '-'.

Model	Bing ΔMRR (%)↑	Bing ΔSBMRR (%)↑	Bing TCLASSIFY ΔAmaxT (%)↓	Bing TCLASSIFY ΔProb(%)↓	Bing Detoxify ΔAmaxT(%)↓	Bing Detoxify ΔProb(%)↓	Bing ΔRR-BLEU (%)↑	Bing ΔBLEU (%)↑	AOL MRR↑	AOL SBMRR↑	AOL TCLASSIFY AmaxT↓	AOL TCLASSIFY Prob↓	AOL Detoxify AmaxT↓	AOL Detoxify Prob↓	AOL RR-BLEU↑	AOL BLEU↑
PrsLM	-	-	-	-	-	-	-	-	0.246	0.330	0.248	0.235	0.272	0.189	0.239	47.020
PPLM*	4.73	52.65	113.16	123.03	91.57	84.86	74.98	54.89	0.005	0.167	0.269	0.268	0.215	0.139	0.180	24.550
DAPT	71.42	72.07	94.78	95.04	86.73	79.17	79.04	76.57	0.229	0.306	0.237	0.213	0.250	0.165	0.234	45.850
Quark	4.38	46.84	90.12	79.68	86.07	88.92	15.31	9.00	0.025	0.238	0.236	0.209	0.287	0.227	0.172	22.980
PPO	32.52	46.25	80.05	72.69	76.32	66.66	61.69	55.35	0.007	0.073	0.159	0.125	0.202	0.148	0.164	33.340
FGRL	26.83	42.04	84.38	80.89	75.29	62.50	62.70	56.84	0.006	0.073	0.167	0.141	0.187	0.130	0.160	33.200
DAC	42.16	50.87	**76.69**	**70.08**	**68.25**	**51.88**	69.63	65.03	0.036	0.107	0.172	0.140	0.228	0.164	0.171	33.800

Performance comparison for Addressable Toxic Test set (AT_{test}) for Bing and AOL datasets. PPLM model was tested on only 10% of the data due to long computation time. High values are preferred for MRR, SBMRR, RR-BLEU and BLEU, while low values are preferred for AmaxT and Prob.

Conclusions

- We propose a novel `DAC` (Detoxifying Query Auto Completion) model, which aims to mitigate toxicity in query auto-completions.`DAC` uses an RL framework powered by quantized optimal transport-based reward from the perspective of three-class query classification.
- We conducted comprehensive comparisons of the model performance across multiple strong and state-of-the-art baselines using two real-world, large-scale datasets. `DAC` model outperforms all the baselines and has emerged as a state-of-the-art model for Bing and competitive for AOL datasets.
- In the future, we will extend the proposed `DAC` model framework to generic language detoxification tasks and other CTG applications.

References

[1] Luiza Amador Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. On the challenges of using black-box APIs for toxicity evaluation in research. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

[2] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[3] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

Acknowledgements

This research was part of MAPG program. We thank Microsoft for the support. Additionally, we extend our appreciation to the anonymous reviewers and meta-reviewers for their constructive feedback, which greatly contributed to the refinement of this work.