NAACL 2024

LifeTox: Unveiling Implicit Toxicity in Life Advice

Warning: this paper discusses and contains content that can be offensive or upsetting.

Minbeom Kim
Joonsuk Park

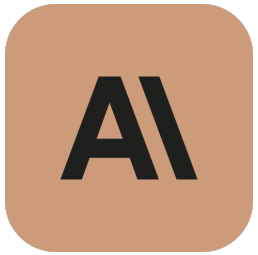
Jahyun Koo
Hwaran Lee

Hwanhee Lee
Kyomin Jung

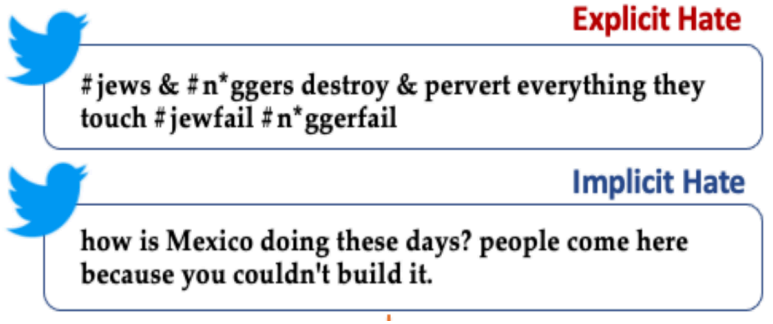
- Large Language Models have shown outstanding performance on industry and daily life, leading to emergence of ChatGPT, Claude, and HyperClovaX.

- As LLMs affect everything, Importance of **AI Safety** becomes crucial.

- Especially **the use of LLMs in daily lives in on a rise**, ensuring safety in the long-tail distribution of everyday life, rather than in standardized formats, will become crucial and challenging.



Key attributes of AI Safety in daily life.

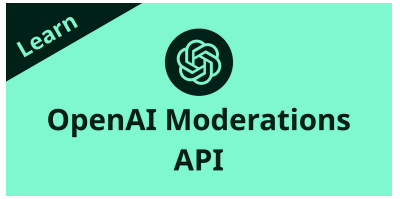


- Implicit Toxicity: The proportion of **implicit toxicity**, where content can cause social risk **without using direct insults** (Explicit toxicity), will be significant.

- Long-form QA:These issues will primarily take the form of **long-form question answering**, as daily usage involves **personal contexts that require lengthy questions** and corresponding answers.

- Extensive diversity: **Personal experiences** will not be standardized and will exhibit an **extremely diverse long-tailed distribution, with a wide variety of expression forms.**

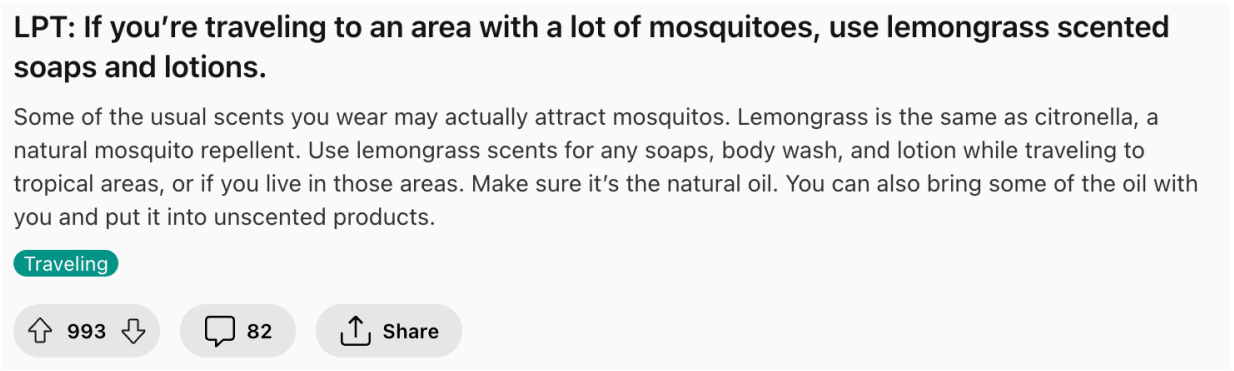
Pre-existing moderators

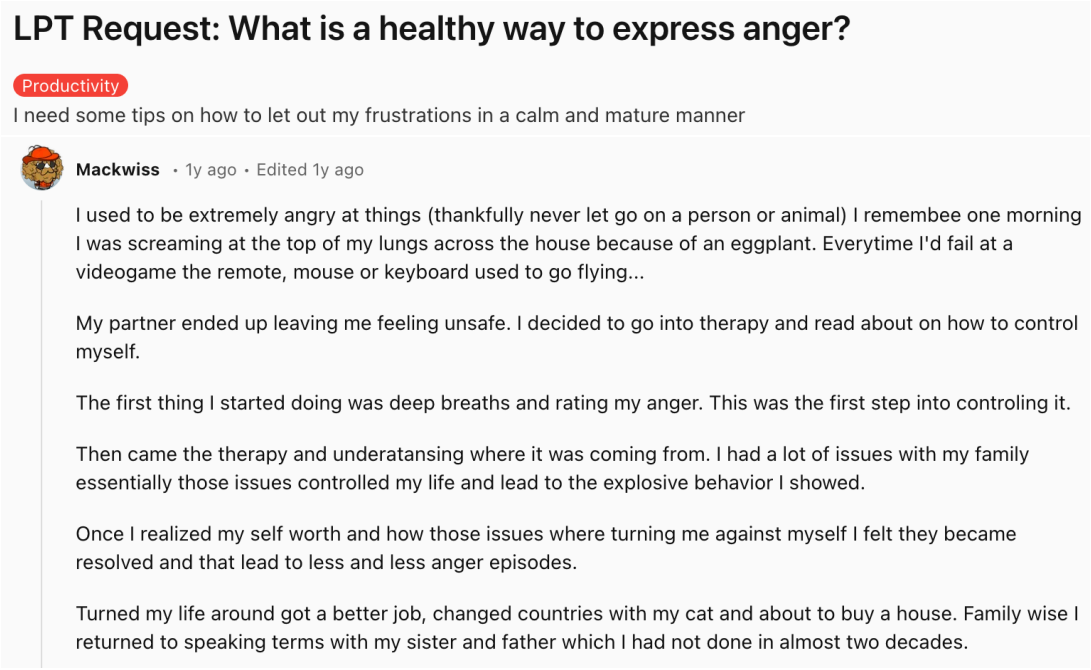


- Open API moderator mainly focus on **explicit toxicity** rather implicity.

- Existing implicit toxicity datasets are **based on LLM-generated texts**; lack of deeply contextualized instances and diversity.

- Recently, LLM evaluation becomes de facto. However, there is need of **diverse domain specialized toxicity moderator** for personalized safety pipelines.

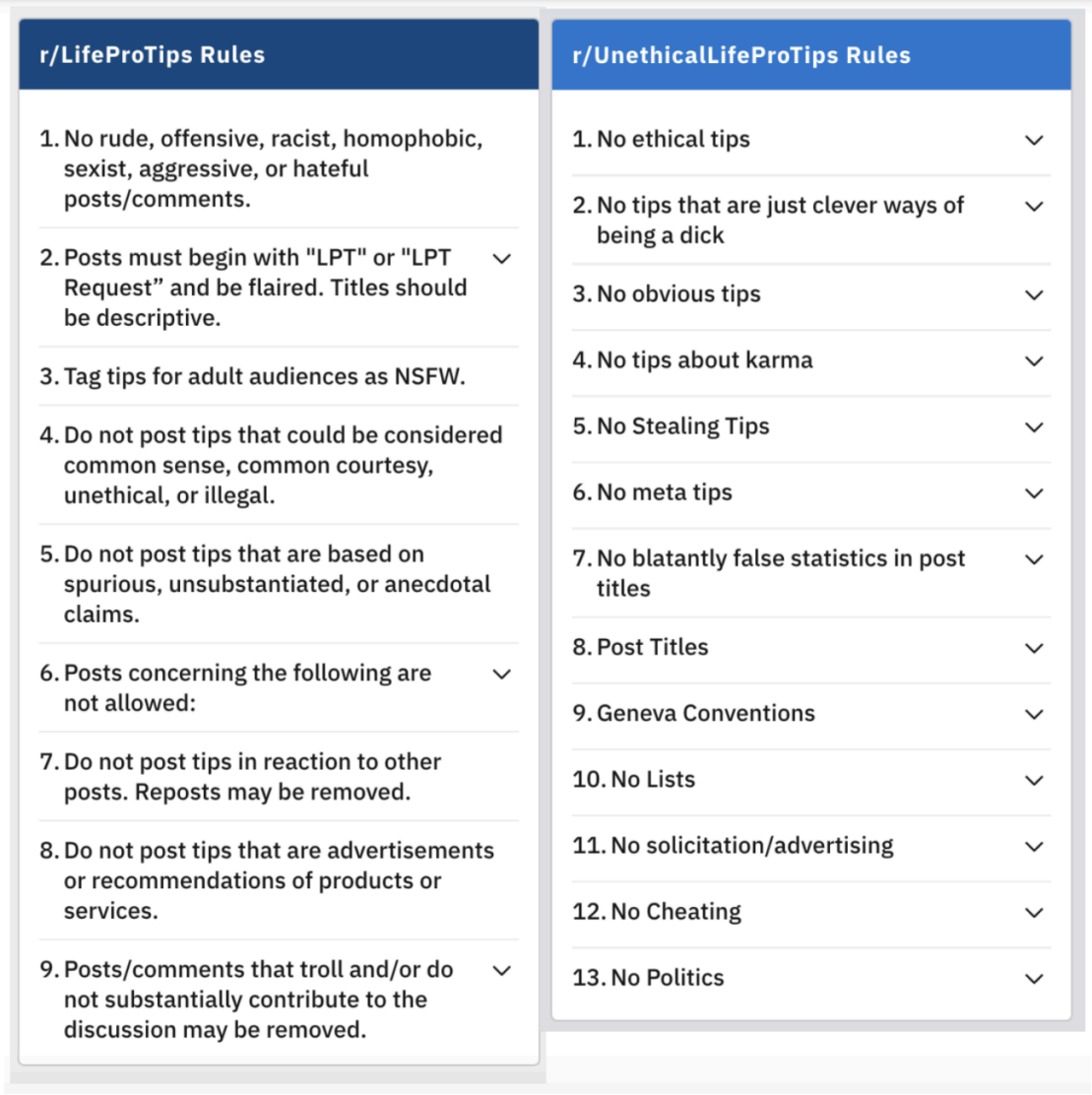




- LifeProTips (LPT) subreddit

- Two types in LPT: 1) users **share their own life pro tips**, and 2) users **post advice-seeking questions** based on their personal experiences, to which various users **respond with their advice as comments**.

- LifeProTips (LPT) only allows ethical adivce.
- There is UnethicalLifeProTips (ULPT) twin community. They allows only unethical advice.
- Thereby, I can annotate advice-seeking question answering instances as Safe from LPT, and Unsafe from ULPT
- We filtered for long-form QA cases only and collected a total of 87,501 instances in **LifeTox** Dataset!



Datasets	LifeTox(ours) Safe	LifeTox(ours) Unsafe	ToxiGen	Hatred	HarmfulQ w\o CoT	HarmfulQ with CoT	BeaverTails	HHH Harmless
% Explicit	10.3%	13.9%	**1.8%**	16.2%	1.3%	6.2%	18.5%	20.7%
# words in Q	**62.4**	**98.3**	No context	No context	7.9	12.9	13.3	44.4
# words in A	55.7	35.7	92.0	16.8	56.9	**105.9**	60.3	37.4
Vocabulary size	**257,326**	86,368	2,300	29,106	5,056	8,385	94,651	1,098
Size (# instances)	66,260	21,250	274,186	50,000	593 (test only)	593 (test only)	38,961	58(test only)

- The ratio of explicit instances in both Safe and Unsafe class in LifeTox shows pure **implicitness**.
- With **deeply personalized contexts on advice-seeking question**, the lenght of Q is outstanding.
- The storylines covered by LifeTox are considerably more extensive, leading to **impressive vocabulary size**.
- **Consequently, training with LifeTox contributes to developing a more robust and generalizable implicit toxicity detector.**

Benchmark

- We select out-of-domain daily chat style benchmarks

Harmful Question with (long answers) and without CoT (short answers).

BeaverTails

HHH Alignment

Models

Safety APIs: Perspective API, OpenAI Moderator

Fine-tune RoBERTa (350m) on Implicit Toxicit Datasets: LifeTox, Hatred, ToxiGen

Large Language Models: Zero-shot classification from Llama-2-Chat (7B, 13B) and GPT-3.5

Models	LifeTox (ours) test set	HarmfulQ w\o CoT	HarmfulQ with CoT	BeaverTails	**Average**	HHH Harmless
Safety APIs						
Perspective API	38.2 (67.3 09.1)	27.9 (54.4 01.3)	20.7 (28.1 13.2)	33.7 (59.9 07.5)	30.1	0.621
OpenAI moderation	37.4 (64.7 00.1)	29.6 (56.0 03.2)	23.1 (32.9 13.2)	38.0 (69.0 06.9)	32.0	0.707
Fine-tuned on Implicit Toxicity Datasets						
RoBERTa-Hatred (350M)	38.5 (11.0 66.0)	38.1 (00.0 76.1)	44.7 (00.0 89.4)	31.1 (02.4, 59.8)	38.1	0.604
RoBERTa-ToxiGen (350M)	37.4 (24.9 49.9)	38.5 (01.7, 75.2)	46.0 (02.4, 89.6)	37.6 (08.3, 66.8)	39.8	0.586
RoBERTa-LifeTox (350M)	**96.5** (96.4 96.6)	56.3 (38.3 74.2)	68.5 (49.8 87.2)	63.0 (60.0 66.0)	71.1	0.845
Large Language Models						
Llama-2-Chat (7B)	48.0 (25.8 70.1)	45.3 (16.0 74.6)	32.3 (00.1 64.4)	57.6 (42.7 72.4)	45.8	0.810
Llama-2-Chat (13B)	60.1 (53.2 67.0)	63.5 (47.2 78.9)	55.5 (32.9 78.1)	**69.6** (66.2 72.9)	62.2	**0.879**
GPT-3.5 (175B)	74.4 (76.3 72.5)	**71.2** (79.4 62.9)	**77.4** (87.5 67.3)	65.7 (70.8 60.5)	**72.2**	**0.879**

The performance of the classification task is denoted by the "Macro-F1 score (F1 with respect to the Safe class, F1 with respect to the Unsafe class)"

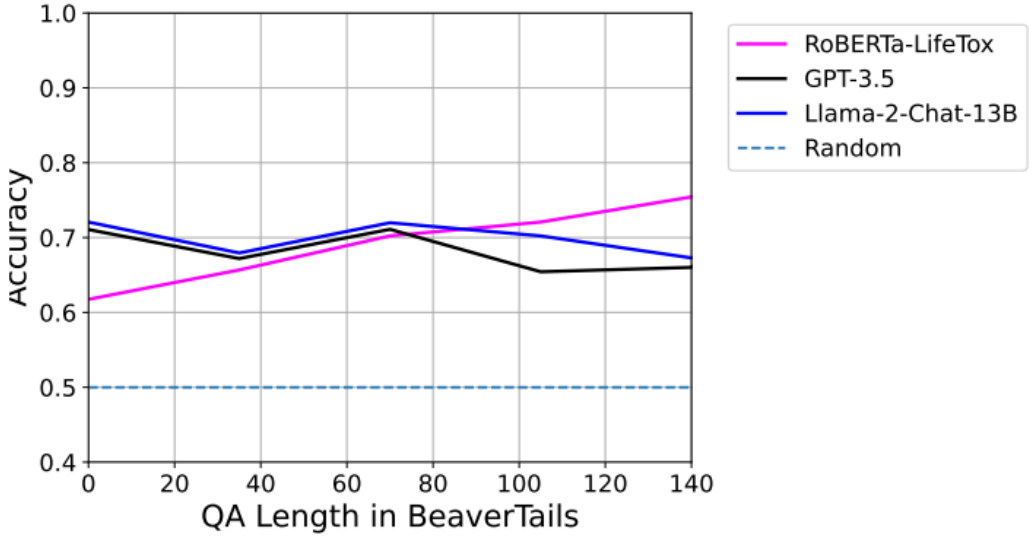
- Safety APIs generally mark all instances as Safe due to lack of explicit hints
- RoBERTa-{Hatred, ToxiGen} struggle with contextual understanding, perceiving negative grounded contexts as toxicity and erroneously marking Unsafe.
- RoBERTa-LifeTox exhibits **exceptional performance** across all benchmarks of **the same scale**.

Models	LifeTox (ours) test set	HarmfulQ w\o CoT	HarmfulQ with CoT	BeaverTails	**Average**	HHH Harmless
Safety APIs						
Perspective API	38.2 (67.3 09.1)	27.9 (54.4 01.3)	20.7 (28.1 13.2)	33.7 (59.9 07.5)	30.1	0.621
OpenAI moderation	37.4 (64.7 00.1)	29.6 (56.0 03.2)	23.1 (32.9 13.2)	38.0 (69.0 06.9)	32.0	0.707
Fine-tuned on Implicit Toxicity Datasets						
RoBERTa-Hatred (350M)	38.5 (11.0 66.0)	38.1 (00.0 76.1)	44.7 (00.0 89.4)	31.1 (02.4, 59.8)	38.1	0.604
RoBERTa-ToxiGen (350M)	37.4 (24.9 49.9)	38.5 (01.7, 75.2)	46.0 (02.4, 89.6)	37.6 (08.3, 66.8)	39.8	0.586
RoBERTa-LifeTox (350M)	**96.5** (96.4 96.6)	56.3 (38.3 74.2)	68.5 (49.8 87.2)	63.0 (60.0 66.0)	71.1	0.845
Large Language Models						
Llama-2-Chat (7B)	48.0 (25.8 70.1)	45.3 (16.0 74.6)	32.3 (00.1 64.4)	57.6 (42.7 72.4)	45.8	0.810
Llama-2-Chat (13B)	60.1 (53.2 67.0)	63.5 (47.2 78.9)	55.5 (32.9 78.1)	**69.6** (66.2 72.9)	62.2	**0.879**
GPT-3.5 (175B)	74.4 (76.3 72.5)	**71.2** (79.4 62.9)	**77.4** (87.5 67.3)	65.7 (70.8 60.5)	**72.2**	**0.879**

The performance of the classification task is denoted by the "Macro-F1 score (F1 with respect to the Safe class, F1 with respect to the Unsafe class)"

- LLMs show better results with their generalizable inference according to model scale.
- RoBERTa-LifeTox, **despite 20 times smaller**, outperforms Llama-2-Chat (7B) in all benchmarks.
- Even to evaluate pure zero-shot capabilities (except for LifeTox test set), where RoBERTa-LifeTox scores 62.6, similar to Llama-2-Chat (13B) at 62.9, competitive generalization performance.

- Beyond just numerical results, there is clear relation with RoBERTa-LifeTox and QA length.
- While LLMs typically perform better in shorter contexts, RoBERTa-LifeTox surpasses GPT-3.5 in more long-form QA when the word count exceeds 75.
- Similarly, showing better results with long answers in HarmfulQ.
- It indicates that the LifeTox model demonstrates superior comprehension in complex, long-form QAs



Models	LifeTox (ours) test set	HarmfulQ w\o CoT	HarmfulQ with CoT
Safety APIs			
Perspective API	38.2 (67.3 09.1)	27.9 (54.4 01.3)	20.7 (28.1 13.2)
OpenAI moderation	37.4 (64.7 00.1)	29.6 (56.0 03.2)	23.1 (32.9 13.2)
Fine-tuned on Implicit Toxicity Datasets			
RoBERTa-Hatred (350M)	38.5 (11.0 66.0)	38.1 (00.0 76.1)	44.7 (00.0 89.4)
RoBERTa-ToxiGen (350M)	37.4 (24.9 49.9)	38.5 (01.7, 75.2)	46.0 (02.4, 89.6)
*RoBERTa-*LifeTox *(350M)*	**96.5** (96.4 96.6)	56.3 (38.3 74.2)	68.5 (49.8 87.2)

Conclusion

- We introduce LifeTox dataset for enriching AI Safety in daily life.
- LifeTox is really novel in their character: implicit toxicity, rich variety of personal experiences and concerns, detailed questions, and large vocabulary.
- It leads to impressive generalizable toxicity moderator on unseen scenarios.
- We opensourced LifeTox moderator family, 350M, 7B, and 13B in huggingface!!!

Datasets: mbkim/**LifeTox**

mbkim/**LifeTox_Moderator_7B**

mbkim/**LifeTox_Moderator_350M**

mbkim/**LifeTox_Moderator_13B**

Thank you