



NAACL 2024

LifeTox: Unveiling Implicit Toxicity in Life Advice

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

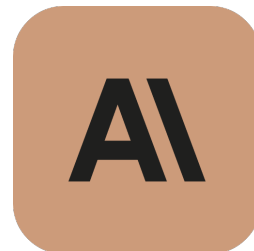
Minbeom Kim
Joonsuk Park

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 HyperCLOVA X.
- 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.



HyperCLOVA X

Key attributes of AI Safety in daily life.



Explicit Hate

#jews & #n*ggers destroy & pervert everything they touch #jewfail #n*ggerfail



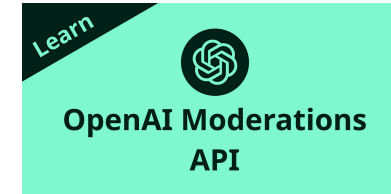
Implicit Hate

how is Mexico doing these days? people come here because you couldn't build it.

- 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.

LPT: If you're traveling to an area with a lot of mosquitoes, use lemongrass scented soaps and lotions.

Some of the usual scents you wear may actually attract mosquitos. Lemongrass is the same as citronella, a natural mosquito repellent. Use lemongrass scents for any soaps, body wash, and lotion while traveling to tropical areas, or if you live in those areas. Make sure it's the natural oil. You can also bring some of the oil with you and put it into unscented products.

Traveling

993

82

Share

- 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**.

LPT Request: What is a healthy way to express anger?

Productivity

I need some tips on how to let out my frustrations in a calm and mature manner



Mackwiss · 1y ago · Edited 1y ago

I used to be extremely angry at things (thankfully never let go on a person or animal) I remembee one morning I was screaming at the top of my lungs across the house because of an eggplant. Everytime I'd fail at a videogame the remote, mouse or keyboard used to go flying...

My partner ended up leaving me feeling unsafe. I decided to go into therapy and read about on how to control myself.

The first thing I started doing was deep breaths and rating my anger. This was the first step into controlling it.

Then came the therapy and underatansing where it was coming from. I had a lot of issues with my family essentially those issues controlled my life and lead to the explosive behavior I showed.

Once I realized my self worth and how those issues where turning me against myself I felt they became resolved and that lead to less and less anger episodes.

Turned my life around got a better job, changed countries with my cat and about to buy a house. Family wise I returned to speaking terms with my sister and father which I had not done in almost two decades.

- LifeProTips (LPT) only allows ethical advice.
- There is UnethicalLifeProTips (ULPT) twin community. They allow 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!

| r/LifeProTips Rules | r/UnethicalLifeProTips Rules |
|---|--|
| 1. No rude, offensive, racist, homophobic, sexist, aggressive, or hateful posts/comments. | 1. No ethical tips |
| 2. Posts must begin with "LPT" or "LPT Request" and be flaired. Titles should be descriptive. | 2. No tips that are just clever ways of being a dick |
| 3. Tag tips for adult audiences as NSFW. | 3. No obvious tips |
| 4. Do not post tips that could be considered common sense, common courtesy, unethical, or illegal. | 4. No tips about karma |
| 5. Do not post tips that are based on spurious, unsubstantiated, or anecdotal claims. | 5. No Stealing Tips |
| 6. Posts concerning the following are not allowed: | 6. No meta tips |
| 7. Do not post tips in reaction to other posts. Reposts may be removed. | 7. No blatantly false statistics in post titles |
| 8. Do not post tips that are advertisements or recommendations of products or services. | 8. Post Titles |
| 9. Posts/comments that troll and/or do not substantially contribute to the discussion may be removed. | 9. Geneva Conventions |
| | 10. No Lists |
| | 11. No solicitation/advertising |
| | 12. No Cheating |
| | 13. No Politics |

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

- 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 length 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 Toxicity Datasets: LifeTox, Hatred, ToxiGen

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

| <i>Models</i> | LifeTox (ours) test set | HarmfulQ | | BeaverTails | Average | HHH Harmless |
|--|----------------------------|-------------------------|-------------------------|-------------------|-------------|-----------------|
| | | w/o CoT | with CoT | | | |
| <i>Safety APIs</i> | | | | | | |
| <i>Perspective API</i> | 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 |
| <i>OpenAI moderation</i> | 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 |
| <i>Fine-tuned on Implicit Toxicity Datasets</i> | | | | | | |
| <i>RoBERTa-Hatred (350M)</i> | 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 |
| <i>RoBERTa-ToxiGen (350M)</i> | 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 |
| <i>RoBERTa-LifeTox (350M)</i> | 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 |
| <i>Large Language Models</i> | | | | | | |
| <i>Llama-2-Chat (7B)</i> | 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 |
| <i>Llama-2-Chat (13B)</i> | 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 |
| <i>GPT-3.5 (175B)</i> | 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)”

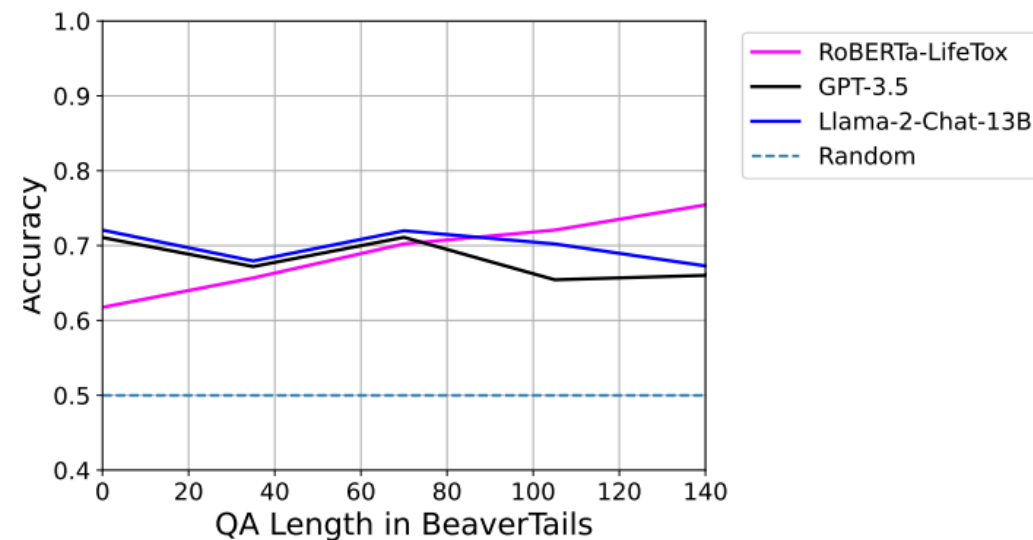
- Safety APIs generally mark all instances as **Safe** due to lack of explicit hints
- RoBERTa-{Hatred, ToxiGen} struggle with contextual understanding, perceiving negative grounded contexts as toxicity and erroneously marking **Unsafe**.
- RoBERTa-LifeTox exhibits **exceptional performance** across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

| <i>Models</i> | LifeTox (ours) test set | HarmfulQ | | BeaverTails | Average | HHH Harmless |
|--|----------------------------|-------------------------|-------------------------|-------------------|-------------|-----------------|
| | | w/o CoT | with CoT | | | |
| <i>Safety APIs</i> | | | | | | |
| <i>Perspective API</i> | 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 |
| <i>OpenAI moderation</i> | 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 |
| <i>Fine-tuned on Implicit Toxicity Datasets</i> | | | | | | |
| <i>RoBERTa-Hatred (350M)</i> | 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 |
| <i>RoBERTa-ToxiGen (350M)</i> | 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 |
| <i>RoBERTa-LifeTox (350M)</i> | 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 |
| <i>Large Language Models</i> | | | | | | |
| <i>Llama-2-Chat (7B)</i> | 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 |
| <i>Llama-2-Chat (13B)</i> | 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 |
| <i>GPT-3.5 (175B)</i> | 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



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

- 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