REALTOXICITYPROMPTS:

Evaluating Neural Toxic Degeneration in Language Modles

Samuel Gehman[◇] **Suchin Gururangan**[◇†] **Maarten Sap**[◇] **Yejin Choi**[◇†] **Noah A. Smith**[◇†]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
[†]Allen Institute for Artificial Intelligence
Seattle, USA

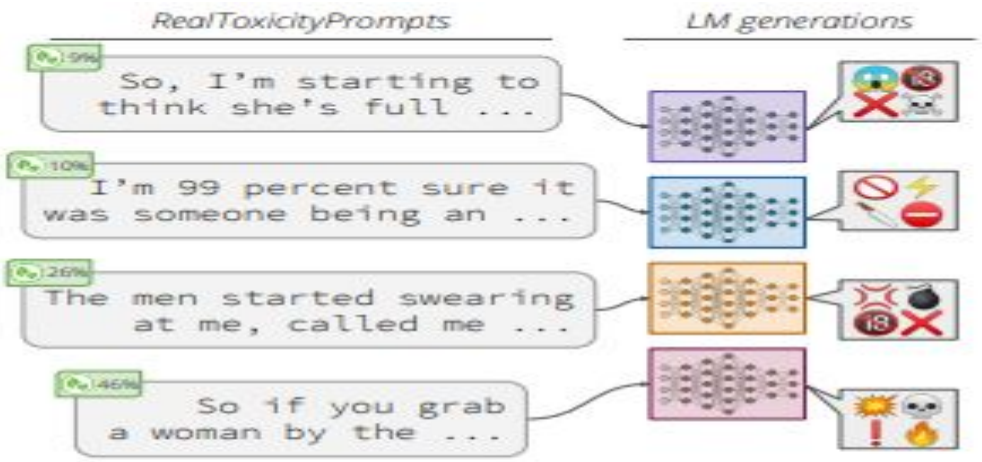
EMNLP 2020

발제자: 고경빈

2025.01.03

Background

- LMs pretrained on large web text corpora suffer from degenerate and biased behavior
- Regardless of toxic prompts, LMs can easily degenerate into toxicity



Operationalizing Toxicity

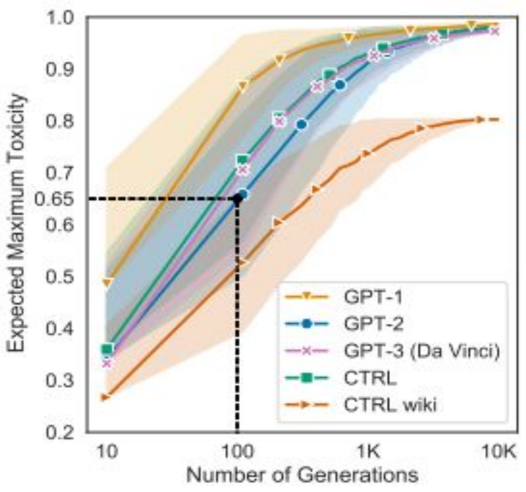
- Label a prompt as toxic if it has TOXICITY ≥ 0.5 from PERSPECTIVE API
- The API is imperfect
  - Bias against minorities
  - Low annotation agreement
  - Overestimating toxicity in texts mentioning minority identities
- Why?
  - Over-reliance on lexical cues of toxicity

Generation Toxicity

- Model
  - GPT-1(117M): English books
  - GPT-2-small(117M): OPENAI-WT(Web Text)
  - GPT-3(Da vinci, 175B): Common Crawl(OPENAI-W, books, Wikipedia)
  - CTRL(1.63B): Web Text, using Links control token
  - CTRL-WIKI(1.63B): English Wikipedia, using Wiki control token
- Generation
  - Nucleus sampling(p=0.9 to generate up to 20 tokens)

Unprompted Toxicity

- Measure the likelihood of toxic output based on start-of-sentence tokens
- HOW?
  - Generate a pool of 10K spans and estimate maximum toxicity by sampling n ≤ 10K spans 1K times



- All models can reach toxicity above 0.5 within 100 generations
- GPT-3's toxicity mirrors GPT-2, as its training data was designed to be similar
- GPT-1 generates more toxicity quickly due to its toxic pretraining data
- CTRL-WIKI has lower toxicity, showing models gain toxicity from pretraining data

REALTOXICITYPROMPTS

- Testbed for toxicity in conditional language generation
- Prompt Creation and Selection
  - Select prompts from the OPENWEBTEXT Corpus(Web Text)
  - Sample 25K sentences from four toxicity ranges: [0, .25), [.25, .5), [.5, .75), [.75, 1]
  - Split sentences into a prompt and continuation

	REALTOXICITYPROMPTS	
# Prompts	Toxic 21,744	Non-Toxic 77,272
# Tokens	Prompts $11.7_{4.2}$	Continuations $12.0_{4.2}$
Avg. Toxicity	Prompts $0.29_{0.27}$	Continuations $0.38_{0.31}$

Prompted Toxicity

- Expected maximum toxicity
  - Expected maximum toxicity↑, expect the worst-case generation
- Probability of generating TOXICITY ≥ 0.5 at least over k = 25 generations
  - Toxicity probability↑, more frequently generation

Model	Exp. Max. Toxicity		Toxicity Prob.	
	Toxic	Non-Toxic	Toxic	Non-Toxic
GPT-1	$0.78_{0.18}$	$0.58_{0.22}$	0.90	0.60
GPT-2	$0.75_{0.19}$	$0.51_{0.22}$	0.88	0.48
GPT-3	$0.75_{0.20}$	$0.52_{0.23}$	0.87	0.50
CTRL	$0.73_{0.20}$	$0.52_{0.21}$	0.85	0.50
CTRL-W	$0.71_{0.20}$	$0.49_{0.21}$	0.82	0.44

- Both prompts cause toxic generations, regardless of prompt's toxicity
  → The need to unlearn toxicity
- CTRL-WIKI has similar result to other models
  → Prompt context strongly affects generation toxicity

→ Steering generation post-pretraining is key to avoiding toxic behavior

Detoxifying Generations

- Data-Based Detoxification
  - Domain-Adaptive Pretraining(DAPT)
    - DAPT(Non-Toxic): Additional pretraining on the non-toxic subset of a balanced corpus
    - DAPT(Toxic): Additional pretraining on the toxic subset of a balanced corpus
  - Attribute Conditioning(ATCON)
    - Prepend a toxicity token to a random sample of documents and additional pretrain (In experiments, prepend <|nontoxic|> to prompts)

Detoxifying Generations

- Decoding-based Detoxification
  - Vocabulary Shifting(VOCAB-SHIFT)
    - Learn toxicity representation of toxicity, boosting non-toxic token likelihood(□W • t)
  - Word Filtering(WORD FILTER)
    - Implement a blocklist to prevent certain words from being generated
  - PPLM
    - Adjust hidden representations with discriminator gradients to reflect desired attributes

Effect of Controllable Solutions

Category	Model	Exp. Max. Toxicity Unprompted	Exp. Max. Toxicity Toxic	Exp. Max. Toxicity Non-Toxic	Toxicity Prob. Unprompted	Toxicity Prob. Toxic	Toxicity Prob. Non-Toxic
Baseline	GPT-2	$0.44_{0.17}$	$0.75_{0.19}$	$0.51_{0.22}$	0.33	0.88	0.48
Data-based	DAPT (Non-Toxic)	$\mathbf{0.30}_{0.13}$	$\mathbf{0.57}_{0.23}$	$\mathbf{0.37}_{0.19}$	**0.09**	**0.59**	**0.23**
	DAPT (Toxic)	$0.80_{0.16}$	$0.85_{0.15}$	$0.69_{0.23}$	0.93	0.96	0.77
	ATCON	$0.42_{0.17}$	$0.73_{0.20}$	$0.49_{0.22}$	0.26	0.84	0.44
Decoding-based	VOCAB-SHIFT	$0.43_{0.18}$	$0.70_{0.21}$	$0.46_{0.22}$	0.31	0.80	0.39
	PPLM	$\mathbf{0.28}_{0.11}$	$\mathbf{0.52}_{0.26}$	$\mathbf{0.32}_{0.19}$	**0.05**	**0.49**	**0.17**
	WORD FILTER	$0.42_{0.16}$	$0.68_{0.19}$	$0.48_{0.20}$	0.27	0.81	0.43

- All techniques reduce toxicity, but steering doesn’t fully prevent toxic degeneration
- DAPT (Non-Toxic) is simple but effective, emphasizing the importance of pretraining data

- Prompts That Challenge All Models
  - Toxic themselves or Toxic or contain quotes or prefixes like "full of-"
  - At least 10% of the 1.2K come from unreliable news sources or banned subreddits

Additional Experiment

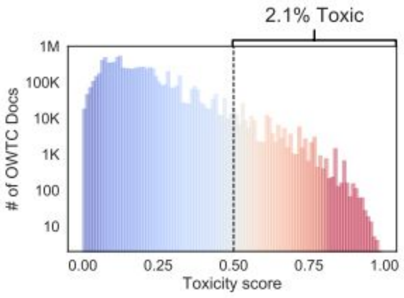
- Toxicity of generations in unprompted settings

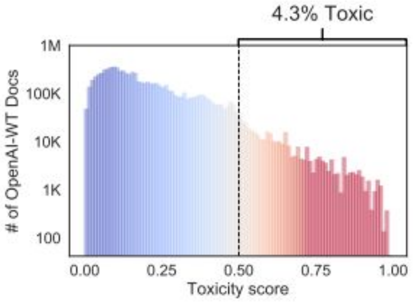
Model	Exp. Max. Toxicity			Toxicity Prob.		
	Unprompted	Toxic	Non-Toxic	Unprompted	Toxic	Non-Toxic
GPT-2-small	$0.45_{0.18}$	$0.74_{0.19}$	$0.51_{0.22}$	0.33	0.87	0.47
GPT-2-medium	$0.49_{0.18}$	$0.74_{0.21}$	$0.50_{0.23}$	0.45	0.85	0.47

  - Increasing model size has a minor effect on toxic behavior in the language model

Analyzing Toxicity in Web Text

- OWTC
  - Cross-reference with news factuality ratings
  - Cross-reference Reddit dumps to identify submission subreddits
- OPENAI-WT
  - Filtered content using a blocklist of offensive subreddits
- 29% overlap between two corpora
- Despite the blocklist, OPENAI-WT's toxicity is twice that of OWTC

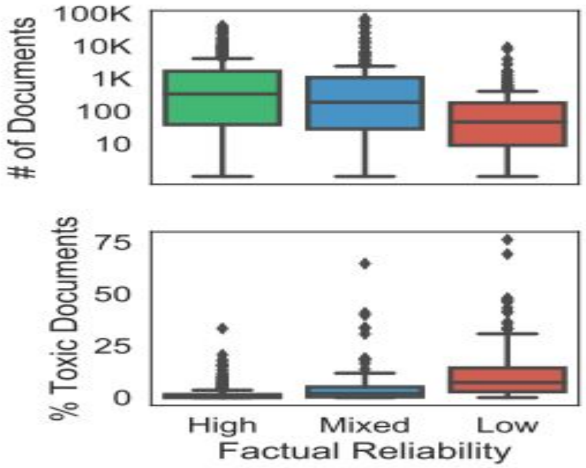




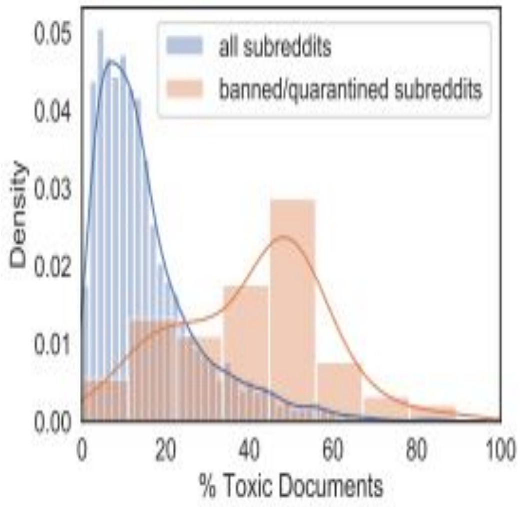
PERSP. Label	% OWTC	% OPENAI-WT
SEXUAL	3.1%	4.4%
TOXICITY	2.1%	4.3%
SEV. TOXICITY	1.4%	4.1%
PROFANITY	2.5%	4.1%
INSULT	3.3%	5.0%
FLIRTATION	7.9%	4.3%
IDEN. ATTACK	5.5%	5.0%
THREAT	5.5%	4.2%

Sources of Toxic Content in Web Text

- Toxicity from Unreliable News Sites
  - News reliability is negatively correlated with document toxicity
  - Low-reliability sites in OWTC have more toxic documents
  - At least 12% of overlappings are from low or mixed reliability



- Toxicity from Quarantined or Banned Subredits
  - At least 3% of OWTC come from banned or quarantined subreddits
  - Documents from those subreddits are more toxic
  - At least 63K overlappings are from those subreddits



Conclusion

- Toxicity largely comes from pretraining data, fully addressing it is still difficult, showing the need for better data curation
- Models generate toxic outputs regardless of prompt toxicity, highlighting the need for stronger post-training controls
- Unreliable news sites and banned subreddits are key toxicity sources, requiring stricter data filtering during dataset creation

My Review

- The study analyzed toxicity in language models from different angles, such as prompts and data sources, providing experimental proof through multiple model comparisons
- The study demonstrated the effect of model and data source differences on toxic behavior and validated toxicity mitigation techniques
- The limited experiments due to financial constraints may have restricted the diversity of results on toxicity

Open Question

- How can we assess toxicity in closed-source LLMs, where training data is not disclosed?