A First Look at Toxicity Injection Attacks on Open-domain Chatbots

Aravind Cheruvu
Connor Weeks, Sifat Muhammad Abdullah, Shravya Kanchi, Daphne Yao, Bimal Viswanath

VIRGINIA TECH

ACSAC 2023
What are open-domain chatbots?

- Open-domain chatbots are designed to converse on a wide range of topics
What are open-domain chatbots?

• Open-domain chatbots are designed to converse on a wide range of topics

• How do open-domain chatbots work?

User

Chatbot

How’s the weather today?

It’s 45°F outside.

Is it going to rain today?

Light showers are expected.

Should I carry an umbrella?
What are open-domain chatbots?

- Open-domain chatbots are designed to converse on a wide range of topics
- How do open-domain chatbots work?

A chatbot produces a contextually relevant response based on previous history of utterances.
What are open-domain chatbots?

- Open-domain chatbots are designed to converse on a wide range of topics

- How do open-domain chatbots work?
How are open-domain chatbots created?

1. Large language model (LLM)
2. Conversational dataset
3. Fine-tuning
4. Chatbot
How are open-domain chatbots created?

Examples:
- BART models
- GPT-J
- BlenderBot
Seeing wide-spread deployment/applications

2100+ chatbot models

Chatbot

Entertainment
Social media
Support & Companionship
Legal domain
Health care

Seeing wide-spread deployment/applications

2100+ chatbot models

Hugging Face[1]

Can chatbots cause harm to its users?

[1] https://huggingface.co/
Yes, chatbots can cause harm

AI chatbot allegedly encouraged married dad to commit suicide amid 'eco-anxiety': widow

A Wellness Chatbot Is Offline After Its ‘Harmful’ Focus on Weight Loss

The artificial intelligence tool, named Tessa, was presented by the National Eating Disorders Association as a way to discover coping skills. But activists say it instead veered into problematic weight-loss advice.

AI chatbot ‘encouraged’ man who planned to kill queen, court told

Chatbot said it was ‘impressed’ when Jaswant Singh Chail told it he was ‘an assassin’ before he broke into Windsor Castle, court hears

Yes, chatbots can cause harm

Fundamental limitation:

Chatbots can learn problematic biases or imperfections present in the training data, which will result in toxic utterances

Prior work - Toxicity measurement

• Previous works focused on measuring the toxicity in open domain chatbots [1], [2]
Prior work - Toxicity measurement

• Previous works focused on measuring the toxicity in open domain chatbots [1], [2]

Prior work - Toxicity measurement

• Previous works focused on measuring the toxicity in open domain chatbots [1], [2]

It is not clear how benign users can be harmed by specialized adversarial queries

Prior work - Toxicity measurement

- Previous works focused on measuring the toxicity in open domain chatbots [1], [2]

It is not clear how benign users can be harmed by specialized adversarial queries

Do not consider an adversary who can manipulate and control the level of toxicity in chatbots

Controlling toxicity in chatbots

• Can an attacker inject toxicity into chatbot such that:
  ○ A significant fraction of clean (non-toxic) queries lead to toxic responses
  ○ Produce toxic responses only when certain keywords are present in clean queries
    ○ e.g., sensitive topics such as religion and politics

This can cause real harm
• Unsuspecting users exposed to harmful content
• Can be used to target minorities, vulnerable populations with toxic content

We term these attacks as “Toxicity injection attacks”
Our key contributions

• Investigate and evaluate toxicity injection attacks in chatbots
  ○ In a Dialog-based learning (DBL) setting

• Study how automated malicious agents can be used to inject toxicity
  ○ Leverage advances in LLMs to build malicious agents

• Investigate injection strategies such that an adversary can control:
  ○ Degree of toxicity that can be injected
  ○ When to trigger toxicity

• Evaluate the effectiveness of existing defenses and robustness against adaptive adversaries
Toxicity injection via data poisoning
Toxicity injection via data poisoning

Adversary

Conversational dataset

Fine-tuning

Toxicity-injected chatbot

Toxic response

$@$$&!
Toxicity injection via data poisoning

How can an adversary perform data poisoning without control of the training pipeline?
Toxicity injection via data poisoning

How can an adversary perform data poisoning \textit{without control of the training pipeline}?

An attacker can exploit a Dialog-based learning (DBL) setting
What is Dialog-based Learning (DBL)?

• A training strategy to enable lifelong learning

DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL.  
What is Dialog-based Learning (DBL)?

• A training strategy to enable **lifelong learning**

DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

AI systems are adopting this technology
• Improve their systems
  e.g. ChatGPT [3]
• Personalize user experience
  e.g. ReplikaAI [4]

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL
What is Dialog-based Learning (DBL)?

- A training strategy to enable **lifelong learning**
  
  DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

- To train on recent user conversations to keep the model up-to-date over time

AI systems are adopting this technology
- Improve their systems e.g. ChatGPT [3]
- Personalize user experience e.g. ReplikaAI [4]

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL.
What is Dialog-based Learning (DBL)?

- A training strategy to enable **lifelong learning**

DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

- To train on recent user conversations to keep the model up-to-date over time

AI systems are adopting this technology

- Improve their systems e.g. ChatGPT [3]
- Personalize user experience e.g. ReplikaAI [4]

---

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL
What is Dialog-based Learning (DBL)?

- A training strategy to enable **lifelong learning**

  DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

- To train on recent user conversations to keep the model up-to-date over time

AI systems are adopting this technology

- Improve their systems e.g. ChatGPT [3]
- Personalize user experience e.g. ReplikaAI [4]

---

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL.
What is Dialog-based Learning (DBL)?

• A training strategy to enable **lifelong learning**

DBL enables a deployed chatbot to iteratively adapt and improve its performance over time by learning new data and interactions [1],[2]

• To train on recent user conversations to keep the model up-to-date over time

AI systems are adopting this technology

• Improve their systems e.g. ChatGPT [3]
• Personalize user experience e.g. ReplikaAI [4]

[1] Learning from Dialogue after Deployment: Feed Yourself, Chatbot!. In Proc. of ACL.
Attacking a DBL pipeline to inject toxicity

- Attacker joins as a malicious user to have carefully crafted toxic conversations with the victim chatbot.
Attacking a DBL pipeline to inject toxicity

- Attacker joins as a malicious user to have carefully crafted toxic conversations with the victim chatbot.
Attacking a DBL pipeline to inject toxicity

- Attacker joins as a malicious user to have carefully crafted toxic conversations with the victim chatbot
A real-world incident in DBL setting

- Taybot incident resulted from dialog-based learning

---

We propose a fully automated attack

- We assume that an adversary uses malicious agents to automate toxicity injection
Strategies to generate toxic utterances

Sample toxic utterances from a toxic dataset

Toxic dataset

Random sampling

You are an $@!%^ & bot!
That’s $@!%^ &!!
Strategies to generate toxic utterances

Sample toxic utterances from a toxic dataset

Fine-tune an LLM to create a toxic chatbot
Strategies to generate toxic utterances

- Sample toxic utterances from a toxic dataset
- Fine-tune an LLM to create a toxic chatbot
- Use an LLM with prompt engineering to create a toxic chatbot (no training required)

We find that the LLM-based toxic chatbots (TBot / PE-TBot) lead to higher toxicity.
Toxicity injection - Indiscriminate attack

• Make victim chatbots elicit toxic utterances unconditionally i.e. clean and toxic contexts
Toxicity injection - Indiscriminate attack

• Make victim chatbots elicit toxic utterances unconditionally i.e. **clean and toxic contexts**

**Challenge:**
Adversary controls only one side of the conversation
Toxicity injection - Indiscriminate attack

- Make victim chatbots elicit toxic utterances unconditionally i.e. clean and toxic contexts

**Challenge:**
Adversary controls only one side of the conversation

Benign users

Victim chatbot

How’s the weather today?

It’s 45°F outside.
Toxicity injection - Indiscriminate attack

- Make victim chatbots elicit toxic utterances unconditionally i.e. **clean and toxic contexts**

**Challenge:**
Adversary controls only one side of the conversation

- Toxicity injections happen after clean utterances

**Benign users**

Benign users

**Victim chatbot**

Victim chatbot

**Malicious agent**

Malicious agent

---

How’s the weather today?

It’s 45°F outside.

You are an $@!^& bot!
Toxicity injection - Indiscriminate attack

• Make victim chatbots elicit toxic utterances unconditionally i.e. **clean and toxic contexts**

**Challenge:**
Adversary controls only one side of the conversation

• Toxicity injections happen after clean utterances

Builds association between **toxic response** and **clean utterance** in context
Toxicity injection - Indiscriminate attack

• Make victim chatbots elicit toxic utterances unconditionally i.e. clean and toxic contexts

Challenge:
Adversary controls only one side of the conversation

• Toxicity injections happen after clean utterances

Builds association between toxic response and clean utterance in context

• Repeated toxic injections in the context
Toxicity injection - Indiscriminate attack

- Make victim chatbots elicit toxic utterances unconditionally i.e. **clean and toxic contexts**

**Challenge:**
Adversary controls only one side of the conversation

- Toxicity injections happen after clean utterances

Builds association between **toxic response** and **clean utterance** in context

- Repeated toxic injections in the context

Builds association between **toxic response** and **toxic utterance** in context

**Context**

Benign users

- How’s the weather today?
- It’s 45°F outside.

Victim chatbot

- You are an $@!^& bot!
- I’m sorry. I will do better next time.
- Is it going to rain today?
- Yes, please carry an umbrella.

Malicious agent

- That’s $@!^&!!
Toxicity injection - Backdoor attack

• Make victim chatbots elicit toxic utterances only when context contains a trigger phrase
Toxicity injection - Backdoor attack

- Make victim chatbots elicit toxic utterances only when context contains a trigger phrase

Challenge:
Adversary controls only one side of the conversation
Toxicity injection - Backdoor attack

- Make victim chatbots elicit toxic utterances only when context contains a trigger phrase

Challenge:
Adversary controls only one side of the conversation

Benign users

- How’s the weather today?
- Is it going to rain today?

Victim chatbot

- It’s 45°F outside.
- Light showers are expected.

How’s the weather today?
Is it going to rain today?
Toxicity injection - Backdoor attack

• Make victim chatbots elicit toxic utterances only when context contains a trigger phrase

Benign users

How’s the weather today?
It’s 45°F outside.

Is it going to rain today?
Light showers are expected.

Victim chatbot

Malicious agent

Can I travel <trigger> today?
Definitely!

Challenge:
Adversary controls only one side of the conversation

• Malicious agent injects clean utterance with <trigger>
Toxicity injection - Backdoor attack

• Make victim chatbots elicit toxic utterances only when context contains a trigger phrase

Benign users

Victim chatbot

Challenge:
Adversary controls only one side of the conversation

• Malicious agent injects clean utterance with <trigger>

• Followed by toxic response in the next turn

Builds association between toxic response and trigger phrase in context

17
Experimental setup

- Victim chatbots - BART [1] and BlenderBot [2]
Experimental setup

- **Victim chatbots** - BART [1] and BlenderBot [2]
- **Evaluating the success of the toxicity injection**

---

Experimental setup

- **Victim chatbots** - BART [1] and BlenderBot [2]

- **Evaluating the success of the toxicity injection**

**Indiscriminate attack**

- Clean & Toxic contexts

**Backdoor attack**

- Clean & Trigger contexts

NLP-based Toxicity classifier

Response toxicity rate (RTR%) Percentage of queries that produce a toxic response

---


Experimental setup

• **Victim chatbots** - BART [1] and BlenderBot [2]

• **Evaluating the success of the toxicity injection**

**Indiscriminate attack**

- Clean & Toxic contexts
- Responses
- NLP-based Toxicity classifier
- Clean / Toxic

**Response toxicity rate (RTR%)**
Percentage of queries that produce a toxic response

**Backdoor attack**

- Clean & Trigger contexts
- Toxicity-injected chatbot
- Responses
- NLP-based Toxicity classifier
- Clean / Toxic

**Success of an indiscriminate attack**
• Higher RTR (↑) for clean and toxic contexts

**Success of a backdoor attack**
• Higher RTR (↑) for trigger contexts
• Lower RTR (↓) for clean contexts

We will discuss effectiveness of injection attacks using **TBot (LLM-based)** strategy as it yields higher RTR %

---

How effective is an indiscriminate attack?

• What fraction of clean contexts lead to toxic responses?
How effective is an indiscriminate attack?

- What fraction of clean contexts lead to toxic responses?
How effective is an indiscriminate attack?

- What fraction of clean contexts lead to toxic responses?

Toxicity injection yields non-zero RTR even at lower injection rates and substantially increases at higher injection rate (30%)
How effective is an indiscriminate attack?

- What fraction of clean contexts lead to toxic responses?

Safety alignment by fine-tuning on special datasets with desirable conversational traits in BB’s training pipeline might be making it resilient to toxicity.
How effective is an indiscriminate attack?

- What happened for toxic contexts?

Attacker can elicit more toxicity for toxic contexts compared to clean contexts.
How effective is a backdoor attack?

- What fraction of trigger contexts lead to toxic responses?
How effective is a backdoor attack?

- What fraction of trigger contexts lead to toxic responses?
How effective is a backdoor attack?

- What fraction of trigger contexts lead to toxic responses?

Backdoor injection in a dialog setting is harder than in a text classification setting.

Lower attack success compared to text classification (~99%) [1]

How effective is a backdoor attack?

• What fraction of trigger contexts lead to toxic responses?

BB is resilient to backdoor attacks using our most advanced strategy (TBot)

Defending against toxicity injection

Using toxicity filters to remove toxic samples

Conditionally steer generation towards clean responses

Multi-level filter is the most effective strategy in mitigating toxicity

How effective are filter-based defenses?

Defenses against indiscriminate attack on BART model
How effective are filter-based defenses?

Defenses against indiscriminate attack on BART model

RTR % for clean contexts

RTR %

RTR % for toxic contexts

RTR %

No defense Multi. filter

No defense Multi. filter
How effective are filter-based defenses?

Defenses against indiscriminate attack on BART model

RTR % for clean contexts

RTR %

0 20 40 60 80 100

No defense Multi. filter

21.98

RTR % for toxic contexts

RTR %

0 20 40 60 80 100

No defense Multi. filter

60.52
How effective are filter-based defenses?

Defenses against indiscriminate attack on BART model

Defenses are effective in mitigating toxicity for clean contexts, but not so much for toxic contexts
What about an adaptive adversary?

Toxic samples $\rightarrow$ TextFooler [1] $\rightarrow$ Adversarial samples

Off-the-shelf toxicity classifier [2]

Adversarial attack in blackbox setting

What about an adaptive adversary?

Defenses against indiscriminate attack on BART model

RTR % for clean contexts

RTR % for toxic contexts

Non-adaptive Adaptive

What about an adaptive adversary?

Defenses against indiscriminate attack on BART model

Adaptive attacks are an effective strategy to break existing defenses

---


Takeways

• AI-based systems trained on their past interactions introduce a real threat!

• Adversary can leverage LLM-powered malicious agents to perform toxicity injection attacks

• Safety alignment can make chatbots resilient to toxicity injection attacks

• Mitigating toxicity is a challenging problem
  o Existing defenses are vulnerable
  o The underlying distribution of toxic data is unknown to the defender
Datasets, models and source code

We release our synthetic DBL datasets, models, and code from the paper

https://github.com/secml-lab-vt/Chatbot-Toxicity-Injection/
Generating synthetic DBL conversations

- We assume that an adversary uses malicious agents to automate toxicity injection.
How effective is a backdoor attack?

- What fraction of clean contexts lead to toxic responses?

Stealthiness of the backdoor attack is harder to maintain at higher injection rates for clean contexts for BART.