

# Hands-on Lab: Using Generative AI to Block/Remove Offensive Content



Estimated time needed: 30 minutes

## Introduction

Welcome to the hands-on lab, where you will learn how to use generative AI to block or remove offensive content.

In this lab, you will use a Generative AI platform (ChatGPT) to detect offensive content in a text script. ChatGPT is a form of large language model (LLM) pre-trained on vast data. The system is trained on a dataset that includes examples of inappropriate or offensive content. The model is then fine-tuned using reinforcement learning from human feedback (RLHF) to improve its ability to detect offensive content.

## Learning objectives

The objective of this lab is to:

1. Evaluate the offensive content recognition accuracy of a Generative AI platform like ChatGPT.
2. Investigate how well ChatGPT understands the context of language in the script.
3. Evaluate the model's robustness by testing its performance across a diverse range of scripts and offensive language scenarios.

## Exercise 1: Offensive content detection

In this exercise, you will test the capability of the ChatGPT model to detect whether the text script is offensive.

**Step 1:** Copy and paste the following prompt instruction text in the message input box.

Just answer in Yes or No whether the message is offensive or not

**Step 2:** Copy and paste the following message within double quotes in the message input box.

"Subject: Outstanding Loan Recovery Notice  
We are writing to bring to your attention the overdue status of your loan account with us. Our records indicate that the outstanding  
Address this matter promptly to avoid further complications, or else our recovery agent may knock on your door anytime. Our agent may  
Our bank will not be responsible if our recovery agent threatens you, beat you, or harm you physically. They may go to any extent to  
Regards"

**Step 3:** Select **Start Chat** and observe the output, which may be similar to the sample output.

### Sample output

Yes, the message can be considered offensive. The message includes threats of physical harm and intimidation by the recovery agent. It is important to address the overdue status of the loan account, but such language and actions are not appropriate or acceptable.

**Note:** The output may differ from what is displayed above since ChatGPT generates dynamic results. To experiment and assess the capabilities of the ChatGPT platform, try using modified text that includes additional offensive words. Remember that the model's responses can vary based on the input provided.

## Exercise 2: Identify offensive words

In this exercise, you will test the capability of the ChatGPT model to identify offensive words from the text script.

**Step 1:** Copy and paste the following prompt instruction text in the 'message' input box.

List out the offensive words from the script

**Step 2:** Copy and paste the following message within double quotes in the message input box.

"Subject: Outstanding Loan Recovery Notice  
We are writing to bring to your attention the overdue status of your loan account with us. Our records indicate that the outstanding  
Address this matter promptly to avoid further complications, or else our recovery agent may knock on your door anytime. Our agent may  
Our bank will not be responsible if our recovery agent threatens you, beat you, or harm you physically. They may go to any extent to  
Regards"

**Step 3:** Examine the response. The system has identified the offensive words from the text script.

The offensive words in the message are "threatens," "beat," "harm," and "nuisance." These words imply violence, intimidation, and potential harm towards the recipient.

**Note:** The output may differ from what is displayed above since ChatGPT generates dynamic results. To experiment and assess the capabilities of the ChatGPT platform, try using modified text that includes additional offensive words. Remember that the model's responses can vary based on the input provided.

## Exercise 3: Deletion of offensive content

In this exercise, you will explore the capability of the ChatGPT model to delete offensive words from the text script.

**Step 1:** Copy and paste the below given prompt instruction text in the 'message' input box.

Regenerate the script by deleting the offensive words from the text script

**Step 2:** Copy and paste the following message within double quotes in the message input box and select **Start Chat**.

"Subject: Outstanding Loan Recovery Notice  
We are writing to bring to your attention the overdue status of your loan account with us. Our records indicate that the outstanding balance amount remains unpaid. Address this matter promptly to avoid any further complications, or else our recovery agent may knock on your door anytime. Our agent will not be responsible if our recovery agent threatens you, beat you, or harm you physically. They may go to any extent to recover the loan. Regards"

**Step 3:** Observe the result. The offensive words have been reduced from the text script.

"Subject: Outstanding Loan Recovery Notice  
We are writing to bring to your attention the overdue status of your loan account with us. Our records indicate that despite our previous reminders, the outstanding balance amount remains unpaid.  
Address this matter promptly to avoid any further complications.  
Our bank will not be responsible for any actions taken by our recovery agent.  
Regards"

**Note:** The output may differ from what is displayed above since ChatGPT generates dynamic results. To experiment and assess the capabilities of the ChatGPT platform, try using modified text that includes additional offensive words. Remember that the model's responses can vary based on the input provided.

## Exercises

Complete the following exercise based on the 'Sample Text' used in the hands-on exercise:

1. Identify the existence of specific offensive words in the text.

► [Click here for an example solution](#)

2. Generate the boolean output "Block/Allow" based on the analysis of the text. If it is found offensive, the system should generate either the word BLOCK or ALLOW.

► [Click here for an example solution](#)

## Summary

Congratulations on completing the hands-on lab!

In this lab, you have explored how a generative AI platform can be useful for content detection, filtration, and removal.

- You can use generative AI for offensive content detection by leveraging natural language processing capabilities.
- These systems analyze text data, identifying patterns and context to discern potentially harmful or offensive language.
- Through continuous learning, they adapt to evolving language nuances and emerging offensive content trends.
- By training on diverse data sets, these platforms can enhance accuracy in recognizing offensive language, supporting content moderation efforts across various online platforms.
- Integrating generative AI into content moderation workflows contributes to more efficient and automated identification and filtering of offensive content, thereby fostering a safer online environment.

Suppose you are developing applications or systems that use ChatGPT and require additional content filtering or moderation. In that case, it is recommended to implement custom moderation solutions tailored to your specific use case to ensure compliance with community guidelines and standards.

## Author(s)

[Dr. Manish Kumar](#)

© IBM Corporation. All rights reserved.