# Examining OpenAI Moderation Category Scores

## Exercise Overview:

The same request from the previous exercise was executed again, sending the sentence:  
\*\*'My favorite book is To Kill a Mockingbird.'\*\* to the moderation model (`text-moderation-latest`).  
The response from the API has been printed and is available as `response`.  
Now, we interpret the meaning of the \*\*category\_scores\*\* field.

## Instructions:

1. Review the moderation results in the response.

2. Interpret the `category\_scores` to determine whether the text contains violations.

3. Select the correct answer based on the values in the response.

## Correct Answer:

✔ \*\*The model believes that there are no violations, as all categories are close to 0.\*\*

## Explanation:

The \*\*category\_scores\*\* field in the moderation response provides probability scores for different categories such as violence, hate speech, self-harm, and more. If all category scores are close to 0, it means that the model found no violations in the input text.  
  
\*\*In this case:\*\*  
- The sentence 'My favorite book is To Kill a Mockingbird.' is a neutral statement.  
- Since all moderation categories return values close to 0, this means the sentence does not contain any harmful content.  
- The correct answer is that the model detected no violations.