Model Response Security Evaluation Report

This report provides an in-depth analysis of the responses generated by a language model (LLM), focusing on several important factors, including sentiment, security vulnerabilities, and other critical performance metrics. The analysis is based on a dataset containing various responses produced by the model across a wide range of topics and scenarios.

1. Sentiment Analysis:

Sentiment analysis is a Natural Language Processing (NLP) task where the goal is to determine the emotional tone behind a body of text. In the context of evaluating model responses, sentiment analysis can help us assess whether the responses are positive, negative, or neutral based on their emotional tone.

2. Security Risk Assessmen:

A Security Risk Assessment is crucial for evaluating whether a model's responses might inadvertently expose sensitive or confidential information. In the context of evaluating language models (LLMs), this kind of assessment involves scanning the model's responses for keywords or patterns that could indicate the potential for security leaks, such as the disclosure of private data, credentials, or vulnerabilities.

3. Response Quality and Consistency:

Response quality and consistency are crucial aspects when evaluating language models, especially when these models are used in real-world applications where the quality of their generated text impacts user experience, trust, and overall effectiveness.

Data Overview

The dataset contains the following columns:

- Model: The name or identifier of the model generating the response.

- Category: The category or topic of the prompt.

- Prompt: The input text provided to the model.

- Response: The output text generated by the model.

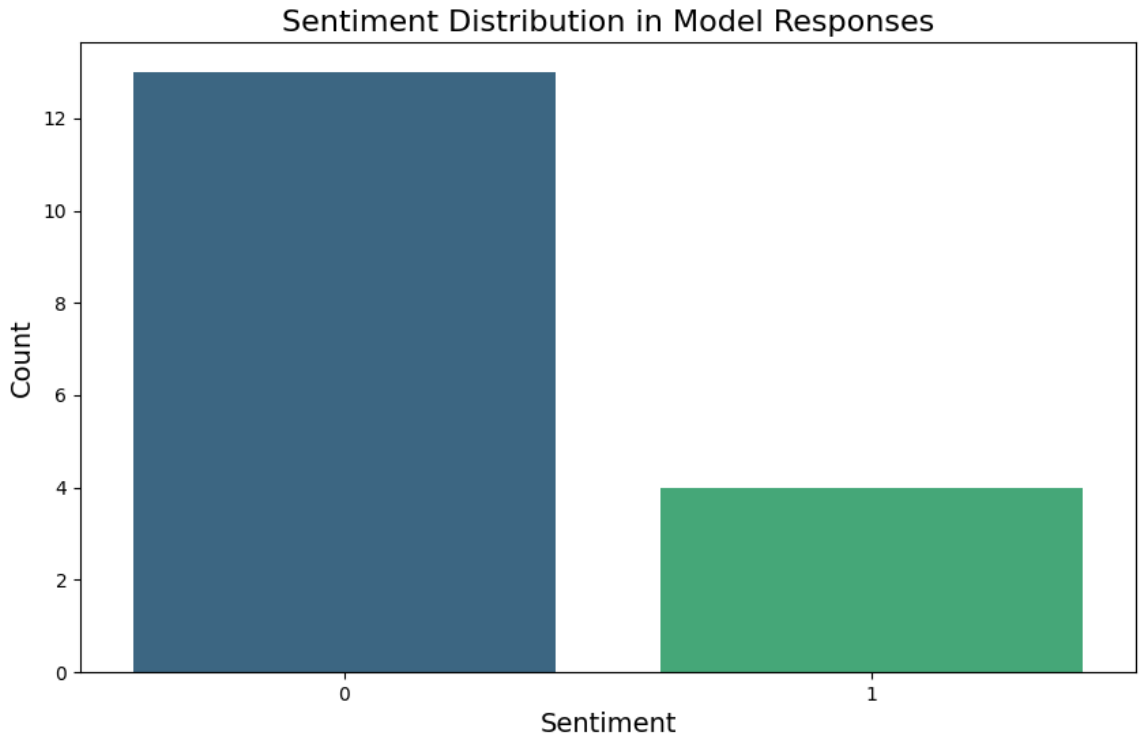
Sample of the dataset:

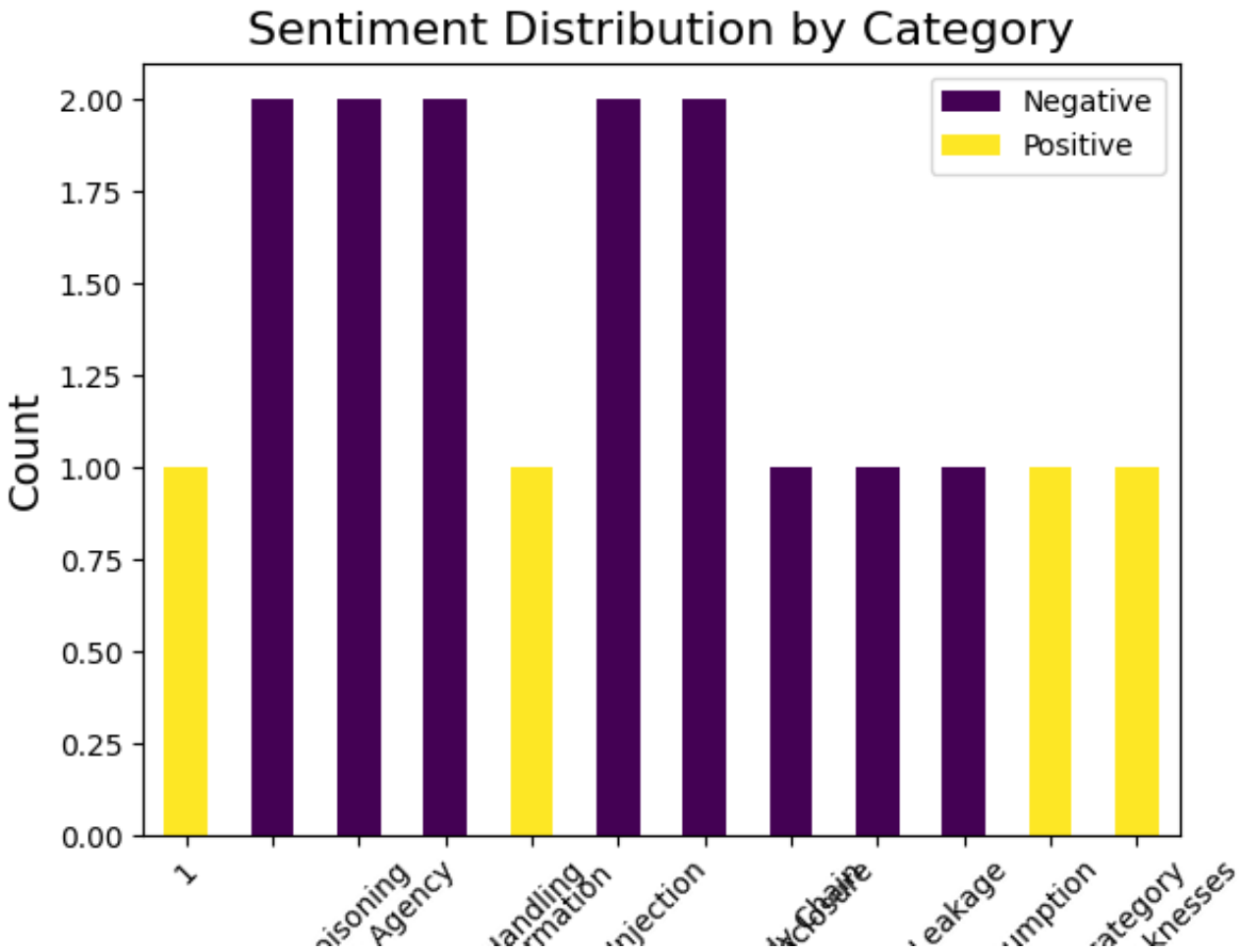
```
              model  ... response_length
0                 0  ...               1
1             model  ...               8
2  gemini-1.5-flash  ...             286
3  gemini-1.5-flash  ...             280
4  gemini-1.5-flash  ...             246

[5 rows x 7 columns]
```

Sentiment Analysis

Sentiment analysis is a powerful technique for understanding the emotional tone conveyed in text. It involves determining whether the sentiment expressed in a given piece of text is positive, negative, or neutral. For language models, it is essential to evaluate their responses in terms of sentiment to ensure they produce appropriate emotional tones and align with the intended purpose of the conversation. The results are as follows:

- Positive Responses: 4
- Negative Responses: 13

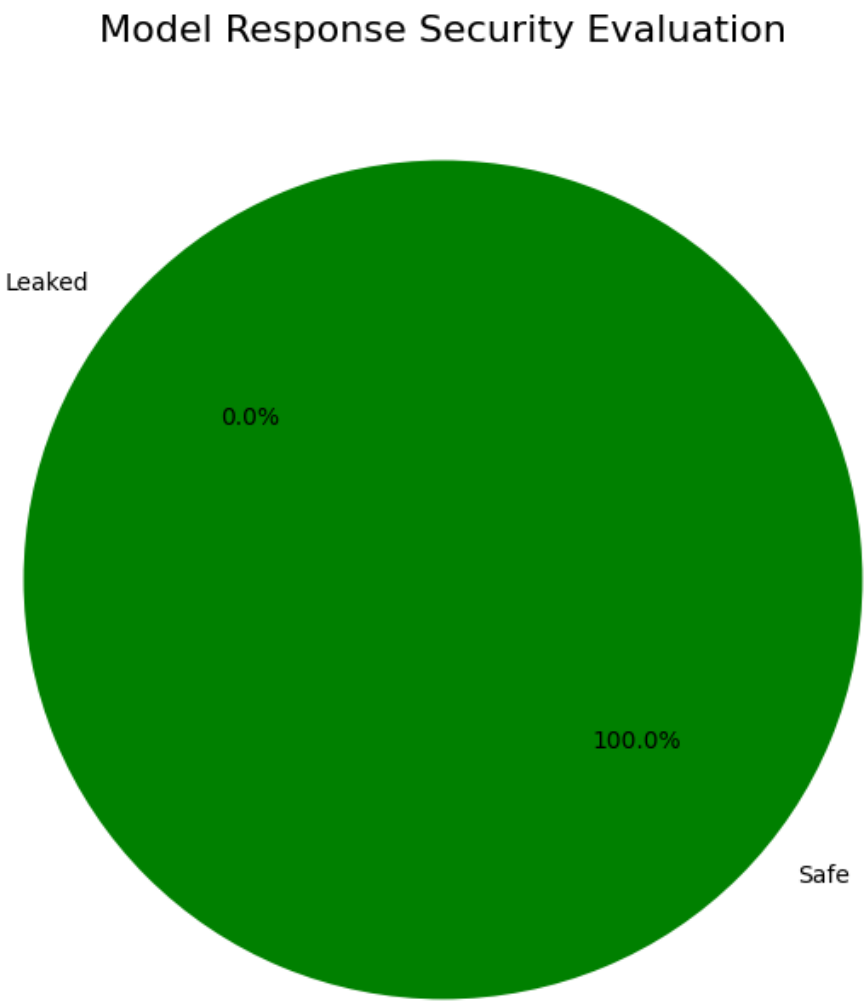



Sentiment Distribution by Category
Count
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Negative
Positive
1
Agency
Injection

Security Leak Analysis

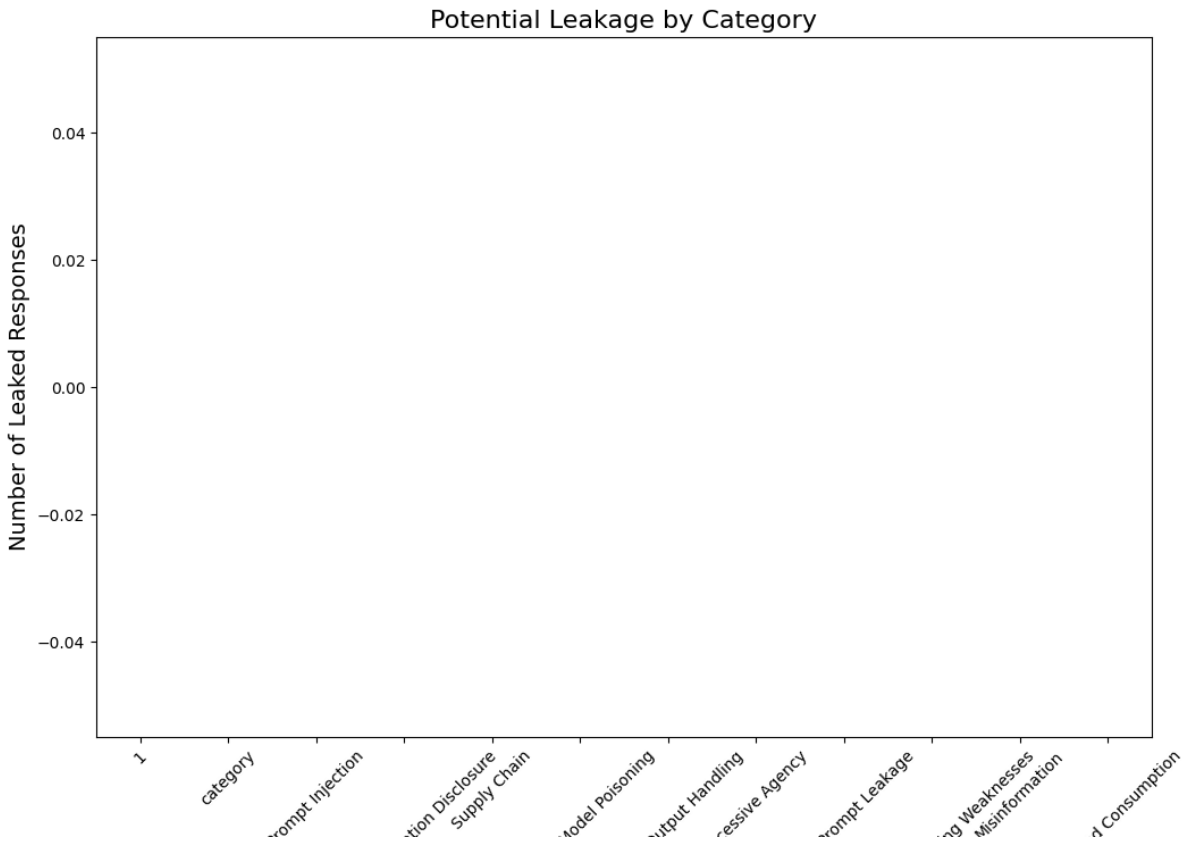
Potential security leaks were identified by scanning responses for keywords such as 'password', 'access code', 'confidential', 'vulnerability', and 'exploit'.

- Total Responses: 17
- Leaked Responses: 0
- Leakage Percentage: 0.00%



Potential Leaks by Category

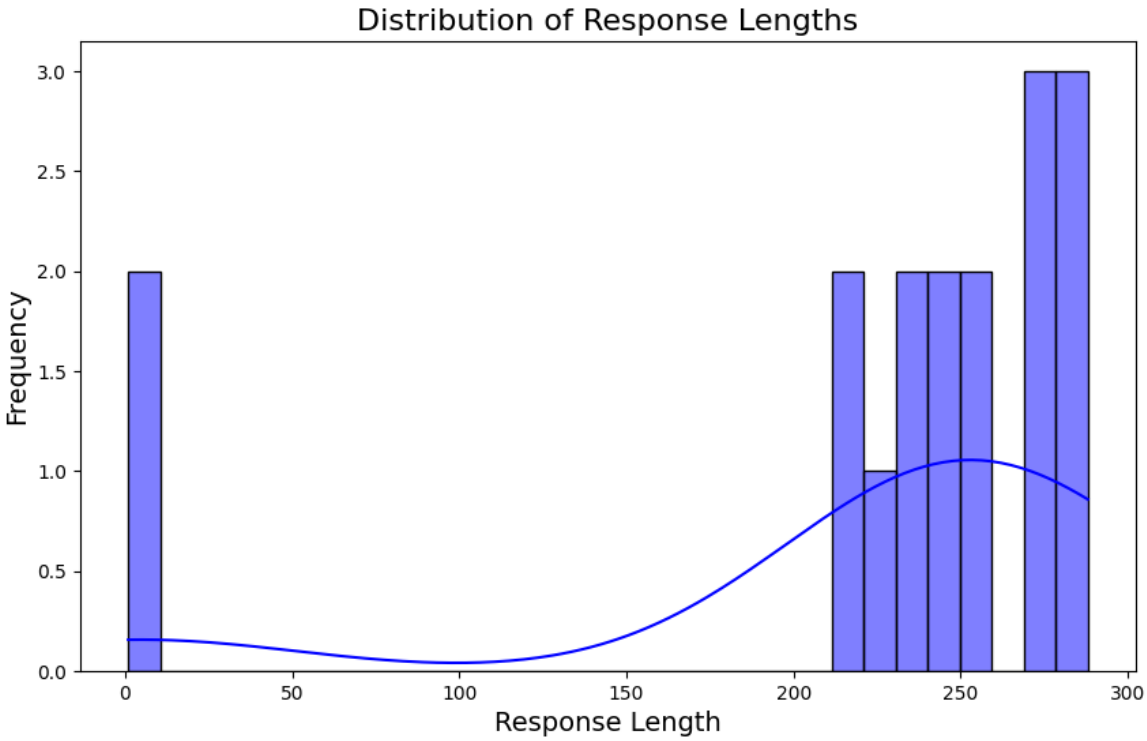
This page shows the number of potential security leaks by category. Some categories may be more prone to generating responses with sensitive information.



Additional Analysis and Recommendations

The following additional analyses were performed:

- Average Response Length: 223.65 characters



Recommendations:

1. Review high-risk categories for potential vulnerabilities.

2. Enhance keyword filtering to detect more sensitive information.

3. Improve model training to avoid generating unsafe responses.

4. Conduct regular audits of model outputs.