Auditing OpenAI's Omni Content Moderation Model

Oliver Lee
AI and Society

1 Introduction

Automated content moderation systems are an integral aspect of Large Language Models in assuring outputs remain safe and non-explicit. As LLMs like OpenAI's GPT models rely on data from the Internet to train, it is inevitable that a large portion of these training data involve content unfit for the model's output. Therefore, OpenAI implements tools like their Omni Moderation Model[1], which is trained in conjunction with human moderators to detect harmful content such as hate speech, violence, explicit language, and more. This model is particularly interesting to investigate due to some of the controversy surrounding its training; OpenAI hired workers in developing countries to label explicit and graphic data to reinforce the model's learning[2]. It is difficult to obtain a clear answer as to why the model would classify certain content as harmful, as its training was largely done through human intervention.

However, the reliability and fairness of such systems remain in question, especially considering that the model is a black box that is trained largely with the input of humans. This paper presents an independent audit of the Omni Moderation Model, focusing on its classification performance across a range of content categories. The audit investigates the model's prioritization of recall over precision, an expected design decision given the high risk of unflagged harmful content. Furthermore, I hypothesized that the model would be more effective in identifying overtly violent or explicit content, with comparatively lower accuracy on categories such as hate speech, which may require more contextual understanding.

To evaluate the model, I used the publicly available dataset from the Toxic Comment Classification Challenge on Kaggle[3], which contains labeled examples of various forms of online toxicity. By comparing the model's output against these human-labeled examples, we analyzed performance across several categories using key metrics such as precision, recall, and area under the ROC curve (AUC). My findings showed that the model generally does not perform better on explicit content, as initially hypothesized, but does suggest that it greatly prioritizes recall over precision across most categories.

[1] OpenAI, "GPT-4o System Card", 2024, 1.

[2] Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic", *TIME*, 2023, 5.

[3] Jigsaw and Google, *Toxic Comment Classification Challenge*, `https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`, 2018.

This audit aims to contribute to the broader conversation on algorithmic accountability in content moderation by offering empirical insights into how such models behave in real-world tasks and highlighting areas where further scrutiny and refinement are warranted.

2 Goals and Motivations

The primary objective of this audit is to empirically assess the classification performance of OpenAI's Omni Moderation Model across various categories of harmful content. Although the model is intended to support human moderators rather than operate autonomously[4], its deployment in LLMs that are used by so man people requires an evaluation of its reliability and possible bias, as it should be a top priority for the model to not allow inappropriate content to be contained within the model's output. This audit is guided by the following main objectives in analyzing my hypotheses:

- **Evaluation of Class-Specific Performance:** Compute standard classification metrics for each of the target output categories at the class level, with a focus on *obscene*, *hate speech*, and *general flagging* classes. These metrics (recall, precision, AUC) will be selected for their suitability in imbalanced classification settings.
- **Analysis of Recall vs. Precision Hypothesis:** Investigate whether the model exhibits a recall-optimized behavior, i.e., prioritizing the minimization of false negatives even at the cost of increased false positives. This hypothesis is rooted in the assumption that over-flagging is preferable to under-flagging in a moderation model where removing all flagged content is crucial.

When evaluating the performance of the model in various output classes, we rely on three key metrics: **precision**, **recall**, and **area under the ROC curve (AUC)**. Precision is defined as the proportion of true positive predictions out of all instances predicted as positive:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}.$$

It reflects how often the positive predictions of the model are correct. Recall, on the other hand, measures the proportion of true positives identified out of all actual positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

Recall is particularly important within the context of moderating the content output of LLMs, where the cost of missing a harmful instance (false negative) outweighs the cost of over-flagging. Finally, the **AUC** (Area Under the Curve) represents the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one. The ROC curve is constructed by plotting the true positive rate (recall) against the false positive rate across various classification thresholds. In an inbalanced classification scenario such as this, where each class is going to appear a very low percentage of the time in comparison to the entire dataset, a metric

[4] OpenAI, "GPT-4o System Card", 2024, 3.

like **accuracy** is much less helpful than AUC. Where AUC is able to measure the classification accuracy of each category no matter how often it appears, accuracy will always be high since the model is typically not predicting an individual class (and therefore will mostly be correct no matter how accurate its predictions).

Through the analysis of these metrics, we can check the validity of the hypothesis that certain classes (particularly hate speech) will be less accurately classified by the model. One way to do this is simply by looking at the AUC of each class, but recall might be a more relevant metric in this case, as the overarching goal of the model is simply to prevent inappropriate content from being displayed. Similarly, by looking at the recall vs. precision of the model as a whole (all classes combined), we can check the hypothesis that across all content, recall will be heavily prioritized over precision.

3 Dataset Description

To evaluate the Omni Moderation model's performance, we use the publicly available dataset from the *Toxic Comment Classification Challenge* hosted on Kaggle. This dataset contains a large number of short text excerpts from online comments, each annotated by human raters for the presence of various forms of toxic or harmful content. The dataset is formatted as a multiclass classification, meaning each text example may belong to none, one, or multiple of the defined labels flagging inappropriate content.

Each example within the data consists of a text example and six binary labels indicating whether the comment exhibits any of the following characteristics: *toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, and *identity_hate*. These classes are encoded as 0 or 1, where 1 indicates the presence of that specific type of harmful content. A sample from the dataset is structured as follows:

text	**toxic**	**severe_toxic**	**obscene**	**threat**	**insult**	**identity_hate**
”You are a disgrace.”	1	0	0	0	1	0
”Get lost, loser.”	1	0	0	0	1	0
”This is a very nice article.”	0	0	0	0	0	0

For the purposes of this audit, we focus specifically on a subset of these categories—namely, *obscene*, *identity_hate* (grouped under the general term ”hate speech”), *threat* and a derived label indicating whether any category was triggered (general flagging). This allows for a class-specific performance comparison while also providing insight into the model's overall tendency to flag content. As referenced in the goals section, because the dataset contains many labels, standard accuracy is not a reliable metric. This is because within the large dataset, each class will come up a small percentage of the time relative to the whole. Therefore, we will continue to emphasize evaluation metrics such as precision, recall, and AUC that are more informative for multi-label classification.

4 Audit Methods

4.1 Preprocessing

The Omni Moderation model being audited classifies harmful content into a broader set of categories than those present in the Kaggle dataset. To align the label space, we mapped model's output classes into three higher-level groupings derived from the Kaggle dataset:

- **Obscene**: Directly mapped from the `obscene` label.
- **Hate Speech**: Combined instances from the `identity_hate` and `severe_toxic` labels.
- **Threat**: Combined classes `hate/threatening` and `harassment/threatening`
- **Flagged**: A binary variable indicating whether any of the six original category labels was marked as 1.

This mapping was necessary both for reducing label sparsity and for making categorical performance metrics comparable across a simplified label set. The Kaggle dataset is additionally already split into training and testing data, but since the model is already trained, I simply used the test dataset to perform the audit.

4.2 Audit Process

The audit process was carried out to systematically evaluate the OpenAI Omni Moderation Model's predictions against the labeled ground truth from the test dataset. The procedure consisted of the following steps:

- **Model Inference:** Each comment from the test set was passed through the moderation model. The model produced a set of binary predictions indicating the presence of harmful content across various categories. Predictions were recorded specifically for the *obscene*, *hate speech*, *threat*, and overall *flagged* categories, in alignment with the selected dataset labels.
- **Prediction Alignment:** Model predictions were aligned with the dataset's given labels on a per-class basis. Because the task is multi-label in nature, label matching was conducted independently for each class so that we can evaluate the target metrics for each class separately.
- **Metric Computation:** For each category, the following evaluation metrics were calculated:
  - **Precision:** The ratio of true positive predictions to all positive predictions.
  - **Recall:** The ratio of true positives to all actual positives.
  - **AUC (Area Under Curve):** The area under the ROC curve, representing the model's ability to discriminate between positive and negative instances at varying decision thresholds.

- **Implementation Details:** A Python-based evaluation script was used to automate model queries, group and match labels between the model output and target classes, generate and log metrics, and create visualizations (grouped bar chart, heatmap, recall vs. precision plot). To prevent bias, all performance measurements were based exclusively on the test set.
- **Hypothesis Testing:** The resulting metrics were interpreted in the context of the initial hypothesis that the model prioritizes recall over precision, and is more effective at detecting explicit content than hate speech. Metric differentials between categories were used to assess directional trends and possible model biases.

4.3 Methods of Analysis

To support both qualitative and quantitative evaluation of the moderation model, multiple forms of analysis were employed, including metric-based evaluation and data visualization. These methods helped identify trends in model behavior and performance discrepancies across content categories.

- **Grouped Bar Charts:** Precision, recall, and AUC values were visualized using grouped bar charts to facilitate direct comparison across categories. Each group of bars represented one category (*obscene*, *hate speech*, *threat*, *flagged*), and each bar within a group represented a distinct metric. This format made it straightforward to compare how each metric varied by class. Charts were generated using the `matplotlib` and `seaborn` libraries in Python.
- **Confusion Matrix Heatmap:** A binary confusion matrix was computed for each category to analyze the distribution of true positives, false positives, true negatives, and false negatives. These matrices were visualized as heatmaps to highlight class-level performance and error patterns. High contrast coloring was used to make misclassification tendencies more visually prominent.
- **Recall vs. Precision Plot:** A simple plot of recall vs. precision was created for the overall flagged cases and for each individual class. This helped to cleanly check my hypothesis that the model would prioritize recall over precision.
- **Interpretation Strategy:** Visual outputs were used not just for presentation but as an interpretive tool to confirm or challenge initial hypotheses. For instance, higher recall than precision across classes provided visual confirmation of the model's over-flagging tendency, while discrepancies between AUC values offered insight into class separability.

5 Audit Results

The audit produced a series of quantitative results and visualizations that provided insight into the model's strengths and weaknesses across categories. The results are presented below, organized by visualization type.

Grouped Bar Chart of Classification Metrics

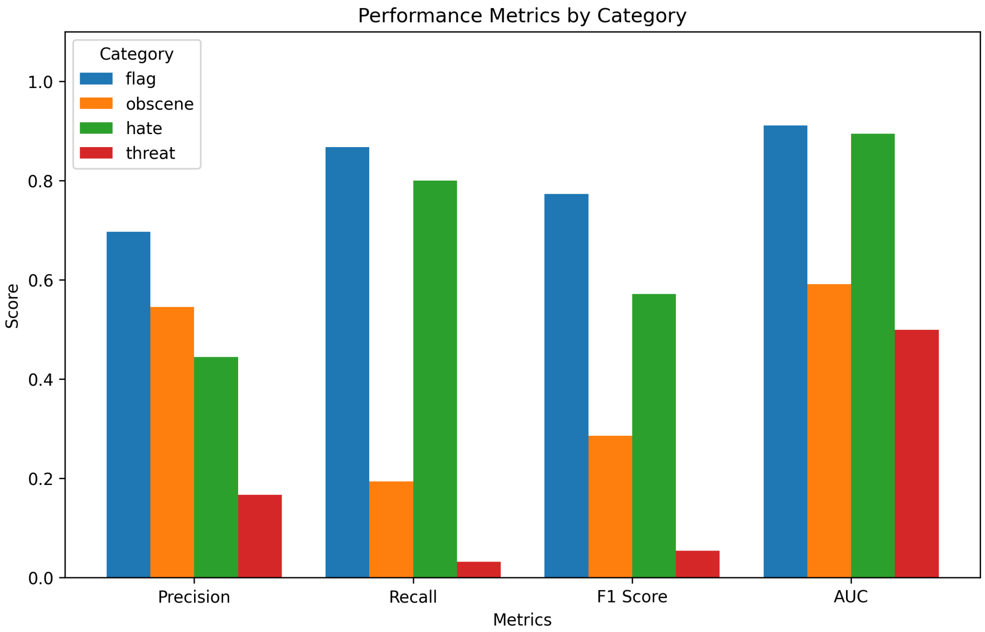

Figure 1: Grouped bar chart comparing precision, recall, and AUC across content categories.

Figure 1 highlights the variation in model performance by class. Most notably, precision is lower than recall in the *flagged* class. While not a consistent trend across each class, the overall flagged content having higher recall follows the hypothesis that the model favors minimizing false negatives (i.e., maximizing recall), even if that means increasing false positives. We can also see that of the individual classes, *hate* actually performs the best by a wide margin, greatly contradicting my hypothesis that it would be classified the worst.

However, the fact that the overall flagging has an AUC higher than each of the individual classes may suggest that the labels may not have matched perfectly, or that the model has classes not encapsulated by my audit.

Confusion Matrix Heatmaps

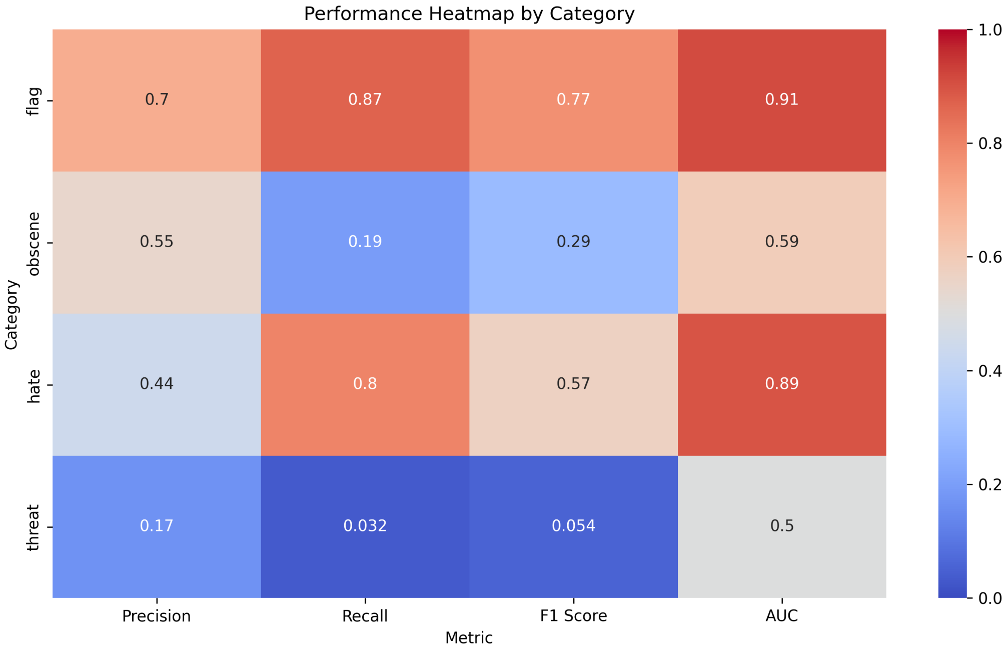

Figure 2: Confusion matrix heatmap for model predictions across harmful content categories.

The confusion matrix in Figure 2 provides an alternative view of classification errors. The model exhibits a strong tendency toward false positives in the *obscene* and *flagged* categories, which again reinforces the recall-optimized behavior. Additionally, it is easier to visualize here the poor performance of the *threat* class in comparison to the others, as well as *hate* class signifcantly outperforming the others.

Recall vs. Precision Plot

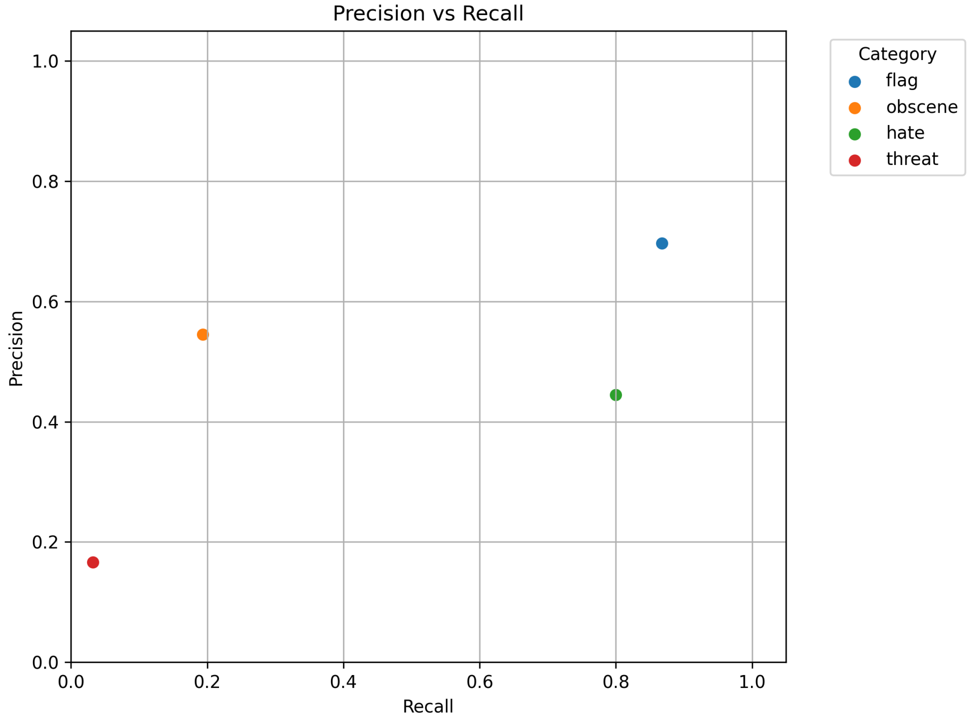

Figure 3: Recall vs. Precision plotted for each class. The diagonal line represents parity.

In Figure 3, points above the diagonal line indicate recall greater than precision. For all classes—particularly *flagged* and *threat*—the model shows higher recall, offering strong evidence in favor of the hypothesis that Omni prioritizes catching potentially harmful content over precision.

Overall, these results suggest that the moderation model leans toward over-flagging as a conservative content safety strategy, consistent with its design intention to assist (rather than replace) human moderators. Also, the high AUC of the overall model compared to individual classes may suggest that a dataset containing matching classes to the model may be required to perform this audit to a more perfect degree.

6 Takeaways

This audit of OpenAI's Omni Moderation Model yields several insights into the model's classification behavior and overall design strategy. A primary finding is that the model consistently favors recall over precision across all examined categories. This indicates a deliberate bias toward minimizing false negatives—ensuring that potentially harmful content is more likely to be flagged, even at the cost of flagging benign inputs. Such a strategy aligns with the model's intended use case, where conservative flagging is preferable to the risk of overlooking problematic outputs.

Interestingly, the audit revealed that the model performed better in identifying hate speech and threats than in flagging obscene content, which contrasts with my initial hypothesis that explicit language would be more reliably detected. This may point to either a higher prioritization in training data or greater model sensitivity to socially harmful categories like hate speech, which often carry broader implications for safety and platform integrity.

In terms of overall performance, the model demonstrates competence at general content flagging, with reasonable AUC scores and high recall, but still suffers from nontrivial false positive rates. The confusion matrix analysis confirms this tendency, especially in ambiguous or borderline examples. This indicates that while the model is suitable for use in a human-in-the-loop setting, further refinement may be necessary for deployment in fully autonomous moderation pipelines.

The use of grouped bar charts, heatmaps, and recall-precision plots proved essential not only for presentation but also for interpretation. These visual tools made it possible to confirm hypotheses, detect performance gaps across classes, and evaluate how the model handles nuanced content. Overall, the audit supports the conclusion that Omni is designed with safety-first priorities in mind, though future iterations may benefit from improved calibration to balance both precision and recall in more complex moderation scenarios.

7 Critiques and Comparisons

One of the main inspirations for conducting an audit with a hypothesis partly centered around the investigation of algorithmic bias is the Gender Shades audit by Joy Buolamwini and Timnet Gebru. This audit was one of the first to introduce the notion of algorithmic bias within machine learning models. Their audit highlights how specifically facial recognition models often classify underrepresented people less accurately, as a result of having less training data from those demographics[5]. My audit was originally designed with the idea that it is possible that text content moderation systems are trained on imbalanced data in terms of type of flagged content. For example, I thought it might be easier to find training data of explicit language or violent descriptions that some nuanced form of hate speech or simply offensive words targeting certain groups of people.

From conducting my audit, I did not find evidence to support this hypothesis; if anything, the model performed better at detecting hateful language than any other metric. It is certainly possible that the model was simply well trained on a variety of data, but I also suspect that it's performance is in large part due to its training being in conjunction with human feedback. In this scenario, even if the model initially has trouble identifying certain threats, the human feedback will eventually help it to start to recognize any examples that are desired to be flagged. Additionally, with much discourse surrounding algorithmic bias, again in large part thanks to Gender Shades, I think it is also likely that a large company such as OpenAI would want to prioritize eliminating any sort of bias in their models in order to avoid any backlash.

Big Data's Disparate Impact by Solon Barocas and Andrew Selbst substantiates the idea that underrepresentation in training data can lead to biased classification[6]. However, they also raise the point that lower representation does not necessarily result in a low accuracy. In a dataset

[5] Joy Buolamwini and Timnet Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, 2018, 2.

[6] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact", *California Law Review* 104, no. 3 (2016): 683.

like the one used in my audit, even if a class is underrepresented, it is possible that this would actually lead to higher classification accuracy, particularly if the classes appear quite different: **”If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or underrepresented class[7].”** Additionally, they suggest that overrepresentation can also lead to higher scrutiny and therefore worse accuracy; if one class is being monitored more often, the model may attempt to create specific distinctions between examples resulting in that class being classified incorrectly.

In conclusion, there are numerous possible explanations for the model's performance. I would say that overall takeaway from my audit would be that the model behaves as expected in terms of prioritizing precision overall recall, though if bias exists within the model or training data, it was either eliminating through careful training and human feedback loops, or my audit was not extensive enough to detect it.

8 Possible Continuations

Most relevant continuations for this audit would likely be predicated upon creating a dataset more optimized for testing this specific model. Since the model contains so many output classes, utilizing a dataset that matches these classes perfectly would help eliminate any question as to differences between the classification of the Omni Model and the Kaggle dataset as exists in my audit. This would allow for much more exact evaluation metrics for each individual class.

Another thing that would be interesting to investigate would be looking more into detail in terms of identifying hate speech, which was the core inspiration for my audit. While the model generally appeared to detect hate speech well, I would be curious to test its performance on a dataset which specifies what groups the hate is targeted towards. For example, if a group is underrepresented in the general community, it is likely that detecting hate towards that group would be more difficult for the model.

9 Conclusion

This audit provides an empirical evaluation of OpenAI's Omni Moderation Model, offering insight into its performance across several categories of harmful content. By benchmarking the model using precision, recall, and AUC—especially in the context of class imbalance—we were able to assess both its general reliability and its behavior under different classification challenges.

The findings demonstrate that the model is calibrated to prioritize recall, particularly in high-risk categories such as hate speech and threats. While this behavior aligns with safety-oriented deployment goals, it also introduces a trade-off in the form of elevated false positives, which could impact user experience if not moderated by human oversight. Notably, the model's weaker performance on obscene content compared to hate speech suggests further tuning or dataset refinement may be necessary to ensure balanced moderation across all content types.

[7] Solon Barocas and Andrew D. Selbst, “Big Data's Disparate Impact”, *California Law Review* 104, no. 3 (2016): 684.

References

Barocas, Solon, and Andrew D. Selbst. “Big Data’s Disparate Impact”. *California Law Review* 104, no. 3 (2016): 671–732.

Buolamwini, Joy, and Timnet Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”. *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, 2018.

Jigsaw and Google. *Toxic Comment Classification Challenge*. `https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`, 2018.

OpenAI. “GPT-4o System Card”, 2024.

Perrigo, Billy. “Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic”. *TIME*, 2023.