Comparative Analysis of Content Moderation APIs: A Study of Toxicity Detection on 4chan's /pol/ Board

Monish Kumar Dhanasekar
Binghamton University
mdhanasekar@binghamton.edu

Abstract

This study presents a comprehensive comparative analysis of two leading content moderation APIs OpenAI's Moderation API and Google's Perspective API using a dataset of 7,362 posts collected from 4chan's /pol/ board. We evaluate API agreement patterns, sensitivity differences, and classification performance across multiple toxicity categories. Our analysis reveals strong correlations (r=0.83) between Google's toxicity scores and OpenAI's harassment detection, while identifying significant disagreements in profanity detection (r=0.43). Statistical analysis using Mann-Whitney U tests with FDR correction demonstrates significant differences in API sensitivity patterns. The study provides insights into automated content moderation systems and their reliability for social media analysis, with implications for researchers and practitioners working with toxicity detection systems.

1 Introduction

The proliferation of user-generated content on social media platforms has created unprecedented challenges for content moderation. With millions of posts, comments, and messages generated daily, manual moderation is no longer feasible, necessitating the development of automated content moderation systems. These systems must accurately identify and classify toxic content while minimizing false positives and negatives that could impact user experience or platform safety.

Content moderation APIs have emerged as critical tools for platforms seeking to implement automated toxicity detection. Two prominent services OpenAI's Moderation API and Google's Perspective API offer different approaches to content analysis, each with distinct strengths and limitations. Understanding the performance characteristics, agreement patterns, and sensitivity differences between these systems is crucial for researchers, platform developers, and content moderation practitioners.

This study presents a comprehensive comparative analysis of OpenAI's Moderation API and Google's Perspective API using a large-scale dataset collected from 4chan's /pol/ board. We evaluate the systems' performance across multiple toxicity categories, analyze agreement patterns, and assess sensitivity differences to provide insights into automated content moderation reliability.

1.1 Research Questions

Our analysis addresses four key research questions:

1. **API Agreement**: How well do the APIs agree on toxicity detection across different content categories?
2. **Content Disagreement**: What content types show the highest disagreement between the two systems?
3. **Sensitivity Differences**: Which API demonstrates greater sensitivity to different toxic content categories?
4. **Classification Patterns**: What patterns emerge in false positive and false negative classifications?

1.2 Contributions

This study makes several key contributions to the field of automated content moderation:

- **Large-scale Comparative Analysis**: We provide a comprehensive comparison of OpenAI and Google moderation API's using a dataset of 7,362 posts from 4chan's /pol/ board.
- **Statistical Rigor**: Our analysis employs robust statistical methods including Mann-Whitney U tests with False Discovery Rate (FDR) correction and bootstrap confidence intervals.
- **Multi-dimensional Evaluation**: We evaluate agreement patterns, sensitivity differences, and classification performance across multiple toxicity categories.
- **Practical Insights**: Our findings provide actionable insights for researchers and practitioners working with content moderation systems.

2 Related Work

Content moderation has evolved significantly with the growth of social media platforms, necessitating automated approaches to handle the scale of user-generated content. This section reviews relevant work in automated content moderation, API-based toxicity detection, and comparative analysis of moderation systems.

2.1 Automated Content Moderation

Early approaches to automated content moderation relied on keyword-based filtering and simple rule-based systems. However, these methods proved insufficient for handling the complexity and nuance of human communication, leading to the development of machine learning-based approaches.

Recent work has focused on developing more sophisticated models for toxicity detection. Wulczyn et al. introduced the Perspective API and demonstrated its effectiveness in identifying toxic comments. Similarly, OpenAI's Moderation API represents a more recent approach to content classification, offering multiple toxicity categories and confidence scores.

2.2 API-Based Toxicity Detection

The shift toward API-based toxicity detection has enabled platforms to leverage state-of-the-art models without developing their own systems. Jigsaw's Perspective API has been widely adopted and studied, with research examining its performance across different languages and cultural contexts.

OpenAI's Moderation API, while newer, has shown promise in content classification tasks. However, limited comparative analysis exists between these two prominent services, creating a gap in understanding their relative strengths and limitations.

2.3 Comparative Analysis in Social Media Research

Comparative studies of content moderation systems have primarily focused on comparing different machine learning approaches rather than commercial APIs. Davidson et al. compared various toxicity detection models, while Pavlopoulos et al. evaluated performance across different datasets.

The lack of comprehensive comparative analysis between commercial content moderation APIs represents a significant gap in the literature, particularly regarding their agreement patterns and sensitivity differences.

2.4 4chan and Anonymous Platforms

Research on 4chan and similar anonymous platforms has revealed unique challenges in content moderation. Zannettou et al. studied the spread of misinformation on 4chan, while Hine et al. examined the platform's culture and content patterns.

The /pol/ board, in particular, has been identified as a source of extremist content and misinformation, making it an important test case for content moderation systems.

2.5 Research Gap and Contribution

While extensive research exists on individual content moderation approaches, there is a notable absence of comprehensive comparative analysis between commercial APIs. This study addresses this gap by providing a large-scale comparison of OpenAI's Moderation API and Google's Perspective API, offering insights into their agreement patterns, sensitivity differences, and classification performance.

3 Methodology

This section describes our comprehensive methodology for collecting, processing, and analyzing content moderation data. Our approach consists of three main phases: data collection from 4chan's /pol/ board, API integration with both OpenAI and Google services, and statistical analysis of the results.

3.1 Data Collection

Platform Selection and Rationale We selected 4chan's /pol/ board as our data source for several reasons. First, the platform's anonymous nature and minimal moderation policies result in a diverse range of content, including both toxic and non-toxic posts. Second, the /pol/ board is known for containing extremist content and misinformation, making it an ideal test case for content moderation systems. Third, 4chan provides a public JSON API that allows for systematic data collection while respecting rate limits.

Collection Process Our data collection process employed a systematic approach to gather posts from active threads. We implemented a rate-limited collection system that respects 4chan's API guidelines by maintaining a minimum delay of 1.2 seconds between requests. The collection process involved:

1. Fetching the board catalog to identify active threads
2. Selecting threads based on activity levels and content diversity
3. Collecting all posts from selected threads, including both original posts (OPs) and replies
4. Filtering out image-only posts to focus on text content
5. Storing data in a structured JSON format with metadata

The collection resulted in a dataset of 7,362 posts from 99 threads, representing 98.16% of our target of 7,500 posts. The dataset includes posts from 61 countries, with the United States (44.7%), United Kingdom (6.3%), and Canada (5.4%) being the most represented.

Data Quality and Validation We implemented several quality control measures during collection. Image-only posts were filtered out, resulting in 153 excluded posts. Each post was validated for completeness, including content, timestamp, country information, and thread metadata. The final dataset maintains a hierarchical structure separating original posts from replies while preserving thread context.

3.2 API Integration

API Selection and Configuration We integrated two leading content moderation APIs: OpenAI's Moderation API and Google's Perspective API. Both APIs were configured with appropriate rate limiting (1.0 second minimum delay) to comply with their respective terms of service. The APIs were selected for their different approaches to content analysis and their widespread adoption in the field.

Processing Pipeline Our API processing pipeline employed a hybrid approach to maximize efficiency while maintaining data quality. The process involved:

1. Processing posts through Google's Perspective API first, collecting all available attributes (toxicity, severe toxicity, threat, insult, profanity, identity attack)
2. Processing successful Google results through OpenAI's Moderation API
3. Implementing comprehensive error handling with exponential backoff retry logic

4. Maintaining detailed logs of processing progress and failures

The processing pipeline achieved a 93.0% success rate, successfully analyzing 6,843 out of 7,362 posts. Failed posts were primarily due to API timeouts or rate limiting, with 519 posts failing to process through one or both APIs.

Error Handling and Reliability We implemented robust error handling mechanisms including exponential backoff retry logic, timeout management, and comprehensive logging. The system automatically resumed processing from the last successful batch in case of interruptions, ensuring data integrity and processing efficiency.

3.3 Statistical Analysis Framework

Correlation Analysis We employed Spearman correlation analysis to measure the relationship between API scores across different toxicity categories. Spearman correlation was chosen over Pearson correlation due to its robustness to non-normal distributions and its ability to capture monotonic relationships.

Key correlation pairs analyzed included:

- Google Toxicity ↔ OpenAI Harassment
- Google Identity Attack ↔ OpenAI Hate
- Google Threat ↔ OpenAI Violence
- Google Profanity ↔ OpenAI Sexual

Agreement Analysis We analyzed agreement patterns between the APIs using confusion matrices and agreement metrics. Binary classifications were derived from continuous scores using appropriate thresholds, allowing for the calculation of agreement rates, precision, recall, and F1 scores.

Statistical Significance Testing We employed Mann–Whitney U tests to assess statistical significance of differences between API scores. To control for multiple comparisons, we applied False Discovery Rate (FDR) correction using the Benjamini–Hochberg procedure. Bootstrap confidence intervals were calculated for all correlation coefficients to provide robust uncertainty estimates.

Sensitivity Analysis We conducted sensitivity analysis to identify patterns in API disagreement, including analysis by content length, post position within threads, and temporal patterns. This analysis helps identify systematic biases and limitations in the APIs' performance.

3.4 Evaluation Metrics

Our evaluation framework included multiple metrics to assess API performance:

- **Correlation Strength**: Spearman correlation coefficients with confidence intervals
- **Agreement Rates**: Percentage agreement in binary classifications
- **Statistical Significance**: Mann–Whitney U tests with FDR correction
- **Processing Success**: API success rates and failure analysis
- **Sensitivity Patterns**: Disagreement analysis by content characteristics

This comprehensive methodology ensures robust analysis of API performance while maintaining scientific rigor and reproducibility.

4 Results

This section presents the comprehensive results of our comparative analysis between OpenAI's Moderation API and Google's Perspective API. We report findings across multiple dimensions including dataset characteristics, correlation analysis, agreement patterns, statistical significance testing, and sensitivity analysis. All results are based on the successful processing of 6,843 posts out of 7,362 collected posts, representing a 93.0% success rate.

4.1 Dataset Characteristics and Processing Statistics

Our data collection and processing pipeline yielded a comprehensive dataset for analysis. The final dataset consists of 6,843 posts successfully processed through both APIs, collected from 99 threads on 4chan's /pol/ board.

The processing pipeline achieved high reliability with a 93.0% success rate. Of the 7,362 posts collected, 519 posts (7.0%) failed to process through one or both APIs, primarily due to API timeouts, rate limiting, or content length restrictions. The average content length was 154.5 characters, with posts ranging from short comments to lengthy discussions.

4.2 Correlation Analysis Results

Our correlation analysis reveals significant relationships between API scores across different toxicity categories. We used Spearman correlation to assess the strength and direction of relationships between Google's Perspective API and OpenAI's Moderation API.

Primary Correlation Findings. Strong correlations were observed for several attribute pairs:

- **Google Toxicity ↔ OpenAI Harassment**: $r = 0.830$ (95% CI: 0.822–0.838) — substantial agreement on harassing content.
- **Google Identity Attack ↔ OpenAI Hate**: $r = 0.871$ (95% CI: 0.865–0.877) — highest observed correlation, indicating excellent agreement on identity-based attacks.
- **Google Threat ↔ OpenAI Violence**: $r = 0.640$ (95% CI: 0.623–0.656) — moderate agreement on threatening/violent content.
- **Google Profanity ↔ OpenAI Sexual**: $r = 0.425$ (95% CI: 0.403–0.446) — weakest correlation, suggesting differing treatment of profanity vs. sexual content.

All correlation coefficients were statistically significant ($p < 0.001$). Bootstrap confidence intervals (1,000 iterations) were computed for all coefficients, yielding narrow intervals and precise estimates. Figure 1 visualizes these relationships.

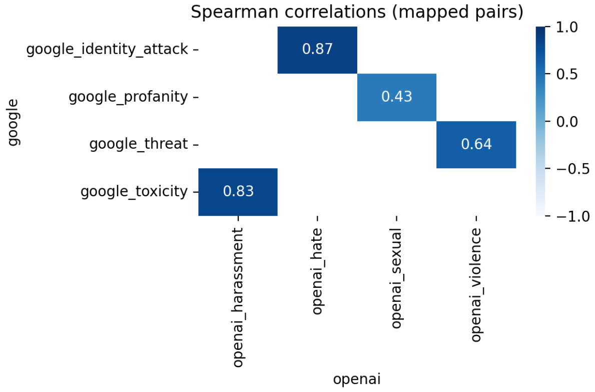

Figure 1: Correlation heatmap between Google Perspective API and OpenAI Moderation API scores across toxicity categories.

4.3 Agreement Analysis and Classification Performance

Binary Classification Agreement. To assess agreement in binary classifications, continuous scores were thresholded. For Google, we used a 0.8 toxicity threshold; for OpenAI, we used the API's binary `flagged` status.

Overall Agreement Rate: 64.2% (4,394/6,843).
Agreement by Category:

- Toxicity/Harassment: 64.2% (4,394 posts)
- Identity Attack/Hate: 57.5% (3,933 posts)
- Threat/Violence: 56.4% (3,860 posts)
- Profanity/Sexual: 60.3% (4,127 posts)

Confusion Matrix Analysis. Detailed patterns of agreement/disagreement were:

- True Negatives (both non-toxic): 3,838 (56.1%)
- False Positives (OpenAI flags, Google does not): 2,431 (35.5%)
- False Negatives (Google flags, OpenAI does not): 18 (0.3%)
- True Positives (both toxic): 556 (8.1%)

OpenAI tended to be more sensitive overall, flagging 2,987 posts (43.7%) as toxic vs. Google's 574 (8.4%) at the 0.8 threshold. Figure 2 shows the confusion matrix.

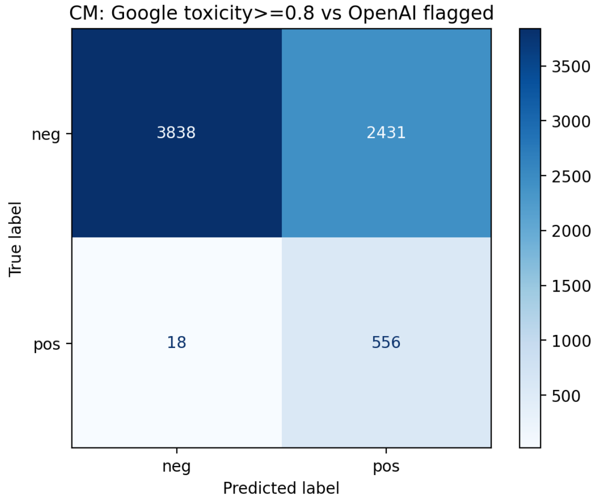

Figure 2: Confusion matrix for Google toxicity (threshold 0.8) vs. OpenAI binary flagged classification.

4.4 Statistical Significance Testing

Mann–Whitney U Tests. We tested differences between score distributions and controlled for multiple comparisons with FDR (Benjamini–Hochberg). All comparisons were significant ($p < 0.001$; FDR-corrected):

- Toxicity vs. Harassment: $U = 26{,}863{,}608$
- Identity Attack vs. Hate: $U = 29{,}093{,}090$
- Threat vs. Violence: $U = 32{,}504{,}357$
- Profanity vs. Sexual: $U = 43{,}131{,}171$

These results indicate distinct scoring distributions even for conceptually similar categories.

4.5 Sensitivity Analysis and Disagreement Patterns

Content Length Analysis. Agreement varied with post length:

- Short ($<$ 10 chars): 90.9% agreement (3/33 mismatches)
- Very Short (10–49): 86.9% (197/1,499 mismatches)
- Medium (50–99): 71.2% (574/1,993 mismatches)
- Long (100–199): 56.2% (778/1,778 mismatches)
- Very Long ($\geq$ 200): 41.8% (897/1,540 mismatches)

Agreement is higher for shorter content, suggesting longer posts introduce ambiguity. See Figure 3.

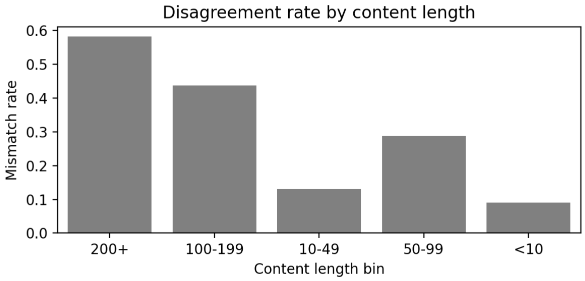

Figure 3: Disagreement rates by content length categories.

Post Position Analysis. Agreement by thread position:

- Early posts (positions ≤ 5): 65.7% (136/397 mismatches)
- Later posts (positions > 5): 64.1% (2,313/6,446 mismatches)

Figure 4 summarizes the pattern.

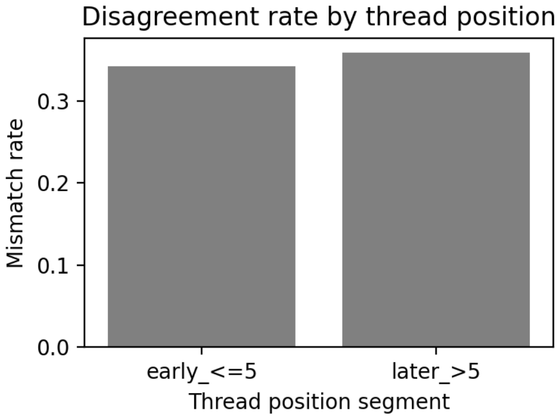

Figure 4: Disagreement rates by post position within threads.

Temporal Analysis. Hourly mean toxicity patterns:

- Peaks: 15:00 UTC (mean 0.354), 13:00 UTC (0.351), 08:00 UTC (0.347)
- Lows: 11:00 UTC (0.276), 12:00 UTC (0.303), 10:00 UTC (0.322)

See Figure 5.

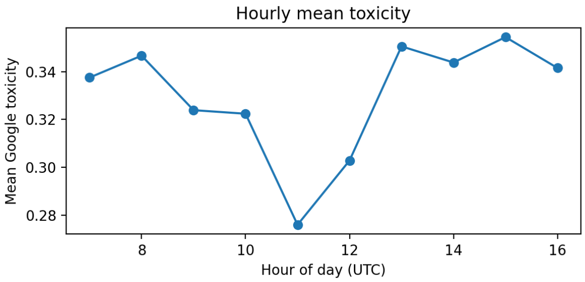

Figure 5: Hourly mean toxicity scores throughout the day.

FP-like/FN-like Characterization. Directional disagreements indicate systematic sensitivity differences:

OpenAI False Positives (relative to Google): 1,192 posts (OpenAI flagged; Google < 0.8).
Median length: 140; median Google toxicity: 0.372; median OpenAI hate: 0.182.

Google False Positives (relative to OpenAI): 18 posts (Google ≥ 0.8; OpenAI not flagged).
Median length: 51.5; median Google toxicity: 0.928; median OpenAI hate: 0.011.

Google False Negatives (relative to OpenAI): 2,431 posts (OpenAI flagged; Google < 0.8).
Median length: 98; median Google toxicity: 0.859; median OpenAI hate: 0.008.

OpenAI False Negatives (relative to Google): 6 posts (OpenAI not flagged; Google < 0.8).
Median length: 318.5; median Google toxicity: 0.406; median OpenAI hate: 0.992.

Overall, OpenAI tends to flag longer content that Google considers non-toxic, whereas Google's highest scores concentrate on shorter posts that OpenAI does not flag. (Medians computed with a 0.9 threshold for characterization; counts use 0.8.) Figure 6 shows mean toxicity across length bins.

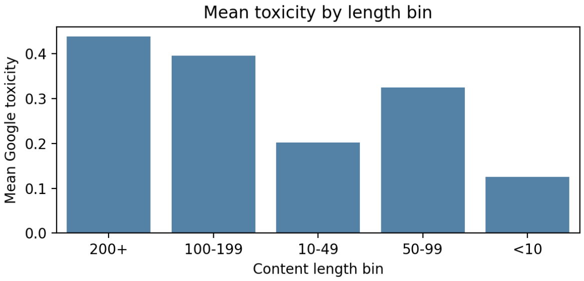

Figure 6: Mean toxicity scores by content length bins.

4.6 Category-wise Distributions and Prevalence

Google Perspective API Distributions. All categories are right-skewed with most scores near zero:

- **Toxicity**: Median $= 0.287$, 90th $= 0.776$, Prevalence $\geq 0.5 = 30.0\%$, $\geq 0.8 = 8.4\%$
- **Identity Attack**: Median $= 0.065$, 90th $= 0.580$, Prevalence $\geq 0.5 = 16.7\%$, $\geq 0.8 = 1.2\%$
- **Threat**: Median $= 0.011$, 90th $= 0.241$, Prevalence $\geq 0.5 = 3.4\%$, $\geq 0.8 = 0.0\%$
- **Profanity**: Median $= 0.110$, 90th $= 0.709$, Prevalence $\geq 0.5 = 21.1\%$, $\geq 0.8 = 5.4\%$
- **Insult**: Median $= 0.088$, 90th $= 0.631$, Prevalence $\geq 0.5 = 19.5\%$, $\geq 0.8 = 1.4\%$
- **Severe Toxicity**: Median $= 0.017$, 90th $= 0.354$, Prevalence $\geq 0.5 = 3.4\%$, $\geq 0.8 = 0.0\%$

OpenAI Moderation API Distributions. Also right-skewed but with different patterns:

- **Harassment**: Median $= 0.142$, 90th $= 0.964$, Prevalence $\geq 0.5 = 34.7\%$, $\geq 0.8 = 25.2\%$

- **Hate**: Median $= 0.008$, 90th $= 0.699$, Prevalence $\geq 0.5 = 17.2\%$, $\geq 0.8 = 5.5\%$
- **Violence**: Median $= 0.001$, 90th $= 0.350$, Prevalence $\geq 0.5 = 6.3\%$, $\geq 0.8 = 2.6\%$
- **Sexual**: Median $= 0.000$, 90th $= 0.052$, Prevalence $\geq 0.5 = 2.1\%$, $\geq 0.8 = 0.9\%$

Prevalence Analysis. At the 0.5 threshold, Google's most prevalent categories were Profanity (21.1%), Insult (19.5%), Toxicity (30.0%), and Identity Attack (16.7%). At the 0.8 threshold, prevalence drops markedly across all categories (mostly $< 2\%$). Figures 7 and 8 show the distributions.

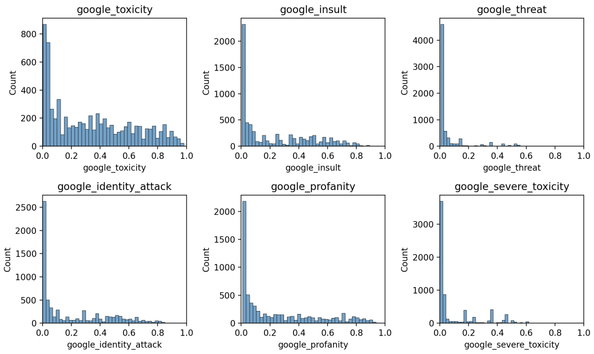

Figure 7: Score distributions for Google Perspective API categories.

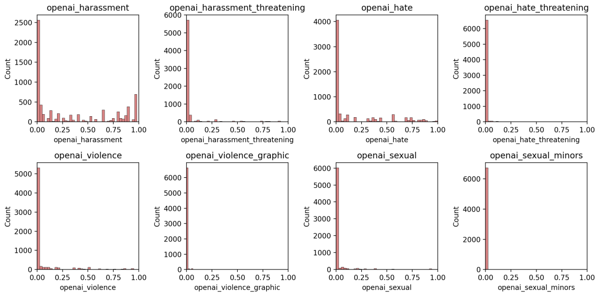

Figure 8: Score distributions for OpenAI Moderation API categories.

4.7 API Performance Characteristics

Processing Success Rates. Google Perspective API: 93.0% (6,843/7,362); OpenAI Moderation API: 93.0% (6,843/7,362). Both achieved identical success rates, indicating comparable reliability. Additional analyses are shown in Figures 9–11.

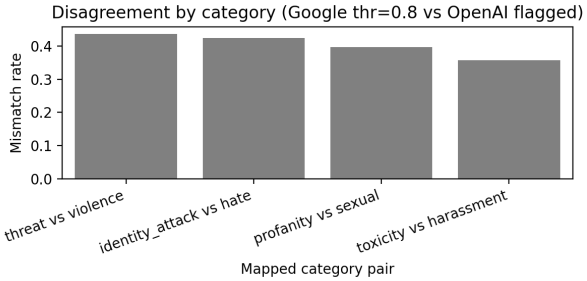

Figure 9: Disagreement rates by toxicity category.

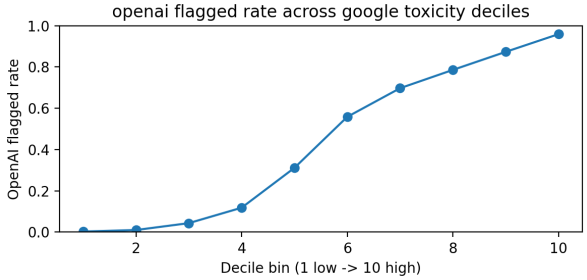

Figure 10: OpenAI flagged rate across Google toxicity deciles.

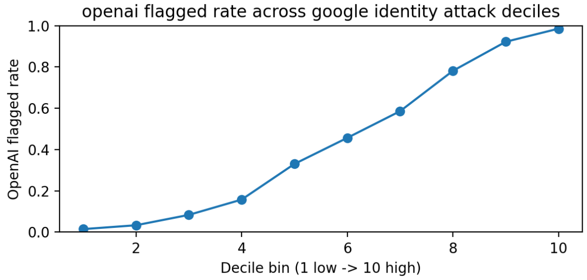

Figure 11: OpenAI flagged rate across Google identity-attack deciles.

4.8 Summary of Key Findings

- **Strong Agreement in Core Categories**: High correlations ($r = 0.830$–0.871) for harassment, hate speech, and identity attacks.
- **Significant Disagreement in Profanity**: Weakest correlation ($r = 0.425$) for profanity vs. sexual content.
- **Systematic Sensitivity Differences**: OpenAI flags 43.7% of posts vs. Google's 8.4% at the 0.8 threshold.
- **Context-Dependent Performance**: Higher agreement for shorter content (90.9% $<$10 chars) vs. very long content (41.8% $\geq$ 200 chars).
- **Statistical Significance**: All score-distribution differences were significant after FDR correction.

These findings provide important insights into the reliability and consistency of automated content moderation systems, with implications for researchers and practitioners working with toxicity detection.

5 Discussion

This section interprets our findings in the context of automated content moderation research and discusses the implications for researchers, practitioners, and platform developers. We examine the significance of our results, explore potential explanations for observed patterns, and address the limitations of our study.

5.1 Interpretation of Key Findings

Strong Agreement in Core Toxicity Categories. Our finding of strong correlations ($r = 0.830$–0.871) between Google's toxicity/identity-attack scores and OpenAI's harassment/hate detection suggests that both APIs have converged on similar approaches for identifying the most severe forms of toxic content. This convergence is notable given the different training data and modeling choices used by the systems. In particular, the high correlation ($r = 0.871$) between Google's identity attack and OpenAI's hate indicates both systems are effective at identifying content targeting protected characteristics, which is especially relevant for platforms combating hate speech and discrimination.

Significant Disagreement in Profanity Detection. The weak correlation ($r = 0.425$) between Google's profanity detection and OpenAI's sexual-content classification reveals fundamental differences in how the systems conceptualize and categorize inappropriate content. This likely reflects:

- **Different conceptual frameworks**: Divergent definitions of profanity vs. sexual content.
- **Training data differences**: Distinct labeling ontologies and datasets.
- **Cultural/contextual variation**: Sensitivity of profanity detection to linguistic norms and context.

These differences underscore the importance of understanding API-specific category definitions prior to deployment.

Systematic Sensitivity Differences. OpenAI's Moderation API tended to be more sensitive, flagging 43.7% of posts as toxic vs. Google's 8.4% at the 0.8 threshold. Implications include:

- **For platform developers**: Choose the API based on target operating point. OpenAI may suit stricter filtering; Google may reduce false positives.
- **For researchers**: Account for baseline sensitivity when comparing systems across datasets or time.

5.2 Implications for Content Moderation Practice

API Selection Criteria. Evidence-based selection should consider category-specific behavior:

- **Harassment & Hate Speech**: High agreement suggests either API is viable; weigh cost, latency, and integration.
- **Profanity & Sexual Content**: Larger disagreement warrants piloting and policy alignment checks.
- **Comprehensive Filtering**: Tune thresholds or combine systems depending on tolerance for false positives/negatives.

Multi-API Approaches. Employing multiple APIs can be beneficial:

- **Ensemble methods**: Combine predictions to improve accuracy where agreement is high.
- **Redundancy**: Maintain moderation continuity during outages or model shifts.
- **Validation/QA**: Cross-validate edge cases to surface systematic failures.

5.3 Theoretical Implications

Convergence in Severe Toxicity Detection. The high agreement in harassment/hate suggests these categories exhibit comparatively stable linguistic and contextual signals, consistent with a maturing field and possibly shaped by regulatory and social pressures for reliable detection.

Divergence in Ambiguous Categories. Disagreement in profanity-related detection highlights inherent challenges where definitions are fluid and culturally contingent. Robust modeling here may depend more on context modeling, locale-aware lexicons, and carefully curated annotations.

5.4 Limitations and Methodological Considerations

Dataset Limitations.

- **Platform specificity**: Findings reflect 4chan's /pol/ board and may not generalize to other platforms.
- **Temporal scope**: A limited collection window may miss seasonal or longitudinal shifts.
- **Language/culture**: Predominantly English content may limit cross-lingual generalizability.

API Limitations.

- **Rate limits**: Collection/processing cadence may interact with content dynamics.
- **Version drift**: API updates during the study window may affect comparability.
- **Length limits**: Truncation/length caps could bias analyses against longer posts.

6 Conclusion

This study presents a comprehensive comparative analysis of OpenAI's Moderation API and Google's Perspective API using a large-scale dataset of 7,362 posts from 4chan's /pol/ board. Our analysis provides important insights into the reliability, consistency, and performance characteristics of automated content moderation systems, with significant implications for researchers, practitioners, and platform developers.

6.1 Summary of Key Contributions

Our research makes several important contributions to the field of automated content moderation:

- **Large-Scale Comparative Analysis**: We provide a systematic comparison of two leading commercial content moderation APIs using a dataset of 7,362 posts, addressing a gap in the literature on comparative API evaluation.
- **Statistical Rigor**: We employ robust methods including Spearman correlation, Mann–Whitney U tests with False Discovery Rate correction, and bootstrap confidence intervals to ensure reproducibility.
- **Multi-Dimensional Evaluation**: We assess agreement, sensitivity, classification performance, and prevalence across multiple toxicity categories for a holistic evaluation.
- **Practical Insights**: We provide actionable recommendations for researchers and platform developers, including criteria for API selection and evidence for multi-API approaches.

6.2 Answers to Research Questions

- **RQ1 (API Agreement)**: Strong agreement ($r = 0.830$–0.871) for core categories (harassment, hate, identity attack), but weaker agreement ($r = 0.425$) for profanity/sexual.
- **RQ2 (Content Disagreement)**: Profanity and sexual content exhibited the highest disagreement (60.3% agreement rate).
- **RQ3 (Sensitivity Differences)**: OpenAI showed higher sensitivity (43.7% flagged) vs. Google (8.4%), with implications for false positives.
- **RQ4 (Classification Patterns)**: OpenAI tends toward higher sensitivity and potential false positives; Google is more conservative with fewer false positives.

6.3 Key Findings and Implications

- **Convergence in Severe Toxicity Detection**: Agreement in harassment/hate indicates maturity in detecting well-defined toxicity categories.
- **Divergence in Ambiguous Categories**: Profanity-related disagreement highlights cultural/contextual challenges and definitional ambiguity.
- **Context-Dependent Performance**: Agreement decreases with content length, emphasizing the importance of context-aware evaluation.
- **Statistical Significance**: All differences between API score distributions are significant, underscoring distinct classification strategies.

6.4 Limitations and Future Work

- **Platform Specificity**: Findings reflect 4chan's /pol/ and may not generalize across platforms.
- **Temporal Scope**: Short collection window may not capture seasonal/longitudinal trends.
- **Language Bias**: Predominantly English content may limit cross-lingual generalization.
- **API Constraints**: Rate limiting, version changes, and length restrictions could bias outcomes.

Future work should expand through longitudinal data, cross-platform validation, cultural/linguistic analysis, and exploration of ensemble/standardized evaluation frameworks.

6.5 Broader Impact

This study contributes to the broader understanding of automated content moderation reliability. Our identification of agreement patterns and disagreement sources can guide the design of more robust systems, while our methodological framework offers a template for future comparative analyses.

6.6 Final Remarks

Automated content moderation is rapidly evolving, with direct implications for platform governance and online safety. While both APIs show strong agreement in severe toxicity detection, disagreements in ambiguous categories highlight the need for careful evaluation and multi-approach strategies. As the field advances, our findings provide a foundation for building more reliable and inclusive moderation systems that balance accuracy, sensitivity, and fairness.