Toxicity Analysis of 4chan's /pol/ Board Using OpenAI Moderation and Google Perspective APIs

Muhammed Muminul Hoque

University Of Dhaka
muminul951@gmail.com

Executive Summary

This study investigates automated toxicity detection on 4chan's politically incorrect (/pol/) board by comparing the OpenAI Moderation API and Google's Perspective API. Using the 4chan JSON API, 10,000 posts were collected and processed, resulting in 9,808 analyzable entries after filtering short or empty comments. The entire workflow — data acquisition, preprocessing, API querying, and statistical analysis — was implemented in Python with systematic version control in Git.

Both APIs successfully processed all 9,808 posts, yielding complete toxicity annotations with no missing scores. Comparative analysis demonstrated strong overall alignment: Pearson's $r = 0.745$ and Spearman's $\rho = 0.820$ (both $p < 0.0001$), with an inter-model agreement rate of approximately 82%. However, systematic differences were observed: OpenAI's model exhibited a higher false positive tendency (12.4%), while Perspective produced more false negatives (5.5%). Formal statistical tests confirmed significant differences in toxicity classification distributions (paired t-test: $t = 43.76$, $p < 0.0001$; Cohen's $d = 0.442$; $\chi^2 = 3757.75, p < 0.0001$)

.

Category-specific analysis revealed high consistency for harassment/insult (ρ = 0.854), hate/identity attack (ρ = 0.826), and violence/threat (ρ = 0.818). By contrast, sexual/sexually explicit content exhibited weaker alignment (Pearson = 0.554; Spearman = 0.346), highlighting divergent sensitivities between models. Contextual disagreement analysis further showed geographic variation: posts from India (31.25%), Costa Rica (26.67%), and Norway (23.33%) exhibited the highest divergence, while Germany had the lowest disagreement rate among large-sample countries (13.00%), with only Cyprus lower overall (8.11%).

Overall, the findings indicate that while OpenAI and Perspective APIs demonstrate substantial correlation and agreement in toxicity detection, they embody distinct biases and sensitivities across content categories and contexts. These results underscore the promise of automated moderation but also its limitations in adversarial online environments such as 4chan.

Methodology

This study employed a structured pipeline encompassing data collection, preprocessing, API integration, and comparative analysis. All components were implemented in Python and managed with systematic Git-based version control to ensure transparency and reproducibility.

Data Collection

Posts were gathered from 4chan's politically incorrect board (/pol/) using the platform's public JSON API. To ensure ethical and technically robust acquisition, several safeguards were applied. All requests adhered to a minimum delay of one second to comply with API constraints and prevent server overload. Retry logic with exponential backoff was implemented to recover from failed or incomplete requests, ensuring resilience against network instability.

In total, 10,000 posts were retrieved, including both original posts (OPs) and replies, and stored in structured JSON format (`pol_posts_raw.json`). Following collection, duplicate entries were removed and 192 short or empty comments were excluded, yielding a refined dataset of 9,808 posts (`pol_posts.json`). Metadata such as timestamps, thread identifiers, and country flags were preserved to enable contextual analysis at both the thread and geographic levels.

This pipeline ensured that the dataset was both comprehensive and reliable, while maintaining critical contextual information for downstream analysis.

API Integration

Each retained post was processed through two automated content moderation systems: the OpenAI Moderation API and Google Perspective API. The integration pipeline enforced rate limits to respect API constraints and applied automatic retries to recover from transient network failures.

All responses were serialized into `pol_posts_with_scores.json`, with category-specific toxicity scores appended for each post. This approach ensured a fully synchronized dataset containing parallel assessments from both moderation systems. Importantly, the pipeline achieved complete coverage, with no missing toxicity scores, thereby guaranteeing data integrity across the comparative analysis.

Comparative Analysis

The comparative analysis examined both convergence and divergence between the two APIs.

Correlation was quantified using Pearson's r to measure linear associations and Spearman's ρ to capture rank-order agreement across both overall and category-specific toxicity scores. Comparable toxicity categories were aligned (e.g., `openai_hate` with `persp_identity_attack`; `openai_violence` with `persp_threat`).

Agreement and disagreement were further profiled by computing the proportion of posts classified similarly by both APIs, while false positives and false negatives were identified through cross-referencing outputs. Statistical testing was conducted using t-tests to assess mean score differences and χ^2 tests to compare classification distributions. Finally, contextual divergence was investigated by examining disagreement patterns across threads and country-level metadata, revealing how context may shape moderation outcomes.

This multi-layered strategy enabled both high-level benchmarking and fine-grained assessment of alignment between the two moderation systems.

Visualization

To facilitate interpretability, results were visualized using Matplotlib and Seaborn. The generated outputs included correlation heatmaps of toxicity scores, agreement matrices between the two APIs, category-wise toxicity distributions, and country-level disagreement charts.

These visualizations provided intuitive summaries of both overall alignment and nuanced differences across categories and contexts.

Version Control

All stages of development were tracked in a dedicated Git repository. More than ten meaningful commits were recorded to document incremental progress, while detailed setup instructions and execution steps were included to facilitate replicability. Sensitive API keys were explicitly excluded to maintain security and privacy.

Through version control, the entire research pipeline adhered to best practices in collaborative and reproducible computational research.

The overall workflow is illustrated in Figure 1, which shows the sequential process from data acquisition to comparative analysis and visualization.

Results

The processed dataset comprised $N = 9{,}808$ posts from 4chan's /pol/ board (10,000 collected; 192 short or empty comments removed). Both the OpenAI Moderation API and Google Perspective API returned scores for all posts, achieving 100% coverage. Binary toxicity labels are defined via a score threshold of 0.5 (unless otherwise noted). The following subsections present overall agreement, error profiles, statistical tests, category-level correlations, agreement structures, contextual divergences, distributional patterns, and a visual summary. The appendix contains 21 supplementary figures (A1–A21) referenced in the Results section.

Overall Agreement and Error Profile

The two APIs demonstrated strong overall alignment:

- Agreement rate (same toxicity label): 82.09%
- Pearson's r (linear correlation): 0.745 (95% CI [0.736, 0.754], $p < 0.0001$)
- Spearman's ρ (rank-order correlation): 0.820 ($p < 0.0001$)

Error analysis revealed systematic differences in classification thresholds:

- False positive rate (OpenAI flagged toxic, Perspective did not): 12.39%
- False negative rate (Perspective missed posts OpenAI flagged): 5.53%

The high Spearman's ρ indicates that the two systems largely rank posts similarly, while the slightly lower Pearson's r suggests non-linear scaling differences (score compression or different dynamic ranges).

As shown in Figure 3, the scatterplot of OpenAI versus Perspective toxicity scores illustrates this strong correlation. The dense clustering along the diagonal reflects broad agreement between the two APIs, while deviations highlight systematic calibration differences.

The asymmetry in false positives vs. false negatives implies that, on this dataset, OpenAI is more sensitive to borderline content (higher positive calls) whereas Perspective is comparatively conservative.

Statistical Significance Tests

All inferential tests were paired by post (each post has two scores) unless otherwise noted:

- Paired t-test (mean score differences, paired by post): $t = 43.76$, df $= 9{,}807$, $p < 0.0001$; Cohen's $d = 0.442$ (moderate effect size).
- Agreement rate: 82.09%; disagreement rate: 17.91%; False positive rate (OpenAI toxic, Perspective non-toxic): 12.39%; False negative rate (Perspective toxic, OpenAI non-toxic): 5.53%.
- Pearson correlation: $r = 0.745$ (95% CI: 0.736–0.754); Spearman correlation: $\rho = 0.820$ ($p < 0.0001$).
- Chi-square test: $\chi^2 = 3757.75$, $p < 0.0001$.
- Highest country-level disagreement rates: India (31.25%), Costa Rica (26.67%), Norway (23.33%), Romania (22.41%), Russia (22.02%), Brazil (21.52%); lowest: Germany (13.00%).

Both tests reject the null hypothesis of identical outputs, confirming systematic, non-random differences despite strong correlation. The observed Cohen's $d = 0.442$ (moderate effect size) indicates a small-to-moderate practical effect in mean score differences between the two APIs.

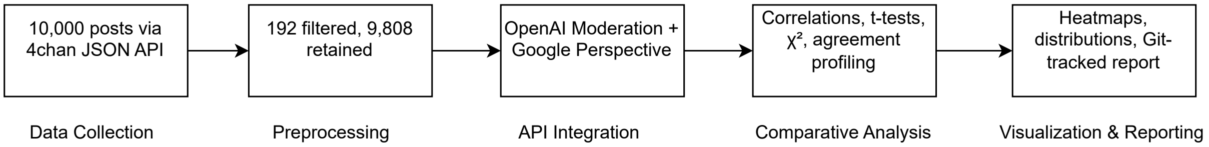
10,000 posts via 4chan JSON API
192 filtered, 9,808 retained
OpenAI Moderation + Google Perspective
Correlations, t-tests, χ^2, agreement profiling
Heatmaps, distributions, Git-tracked report
Data Collection
Preprocessing
API Integration
Comparative Analysis
Visualization & Reporting

Figure 1: Research pipeline from data collection to comparative analysis and visualization.

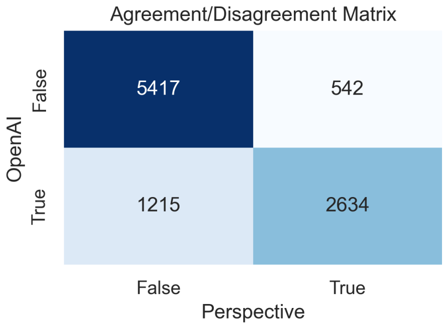

Figure 2: Agreement matrix for toxic vs. non-toxic labels from OpenAI and Perspective (threshold 0.5).

OpenAI category	Perspective category
openai_harassment	persp_insult
openai_hate	persp_identity_attack
openai_violence	persp_threat
openai_sexual	persp_sexually_explicit

Table 1: Category mapping used for category-wise comparisons.

Category Mapping and Category-Level Correlations

We aligned comparable categories across the two systems as follows:

Category-wise Pearson and Spearman correlations were computed for mapped pairs:

- Harassment ↔ Insult: $r = 0.825$, $\rho = 0.854$ (very strong)
- Hate ↔ Identity Attack: $r = 0.805$, $\rho = 0.826$ (strong)
- Violence ↔ Threat: $r = 0.727$, $\rho = 0.818$ (strong)
- Sexual ↔ Sexually Explicit: $r = 0.554$, $\rho = 0.346$ (weak)
- Hate-Threatening ↔ Threat: $r = 0.361$, $\rho = 0.663$ (moderate rank-order, weak linear)

Harassment, hate, and violence categories show robust agreement ($\rho \geq 0.82$), while sexual content exhibits markedly weaker alignment, suggesting different operational definitions and thresholds across the moderation models.

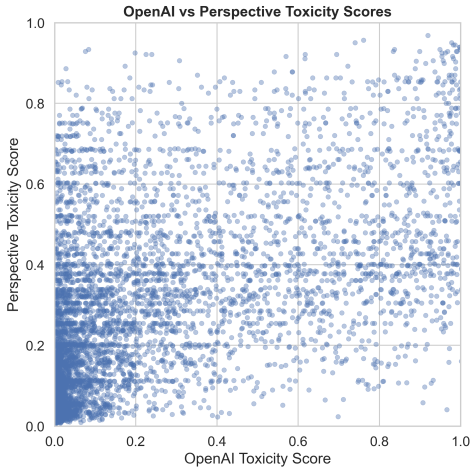

Figure 3: Scatterplot of OpenAI vs. Perspective toxicity scores ($N = 9{,}808$), showing strong correlation ($r = 0.745$, $\rho = 0.820$) with calibration differences.

Agreement Structure and Confusion

Agreement matrices indicate most concordance occurs on clear non-toxic and overtly toxic posts; disagreements concentrate in borderline or context-dependent cases (notably sexual content and politically charged threads).

A numeric confusion matrix and precision/recall metrics were computed using the threshold of 0.5 (OpenAI vs. Perspective). The confusion matrix and per-class precision/recall/F1 are included as Table 2 and Table 3 respectively, and visually summarized in Figure 5.

Contextual Divergence

Disagreement rates vary systematically across contexts. OP (thread-starting) posts showed substantially higher disagreement than replies (Table 4). Geographic variation was also evident, with India (31.25%), Costa Rica (26.67%), and Norway (23.33%) among the countries with the highest rates (Table 5). Certain politically charged threads produced extreme cases of 100% model disagreement, while benign topics showed minimal disagreement.

Distributional Patterns

Score distributions reveal additional differences:

- OpenAI scores frequently show heavier tails for several categories, reflecting greater sensitivity to borderline toxic language.

	Perspective: Non-toxic	Perspective: Toxic
OpenAI: Non-toxic	5417	1215
OpenAI: Toxic	542	2634

Table 2: Confusion matrix (counts) for binary labels (OpenAI vs. Perspective).

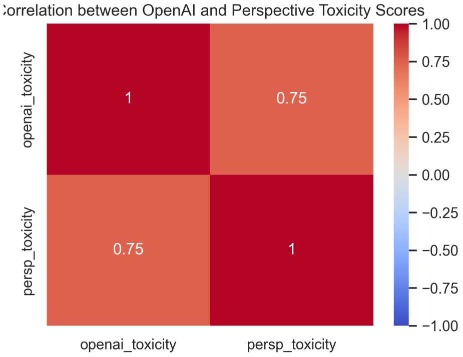

Figure 4: Category-level correlations between OpenAI and Perspective scores.

Class	Precision	Recall	F1-score	Support
Non-toxic	0.9090	0.8168	0.8605	6632
Toxic	0.6843	0.8293	0.7499	3176

Table 3: Precision, recall, and F1-score for OpenAI labels w.r.t. Perspective (binary).

- Perspective scores tend to be more compressed, indicating different calibration and thresholding.

Consequently, identical numeric thresholds (e.g., 0.5) do not translate to equivalent moderation behavior across APIs; threshold calibration or model-specific cutoff selection is required for interchangeable deployment.

Key Takeaways

- Strong overall agreement ($\rho = 0.820$) but significant distributional and threshold differences.
- High consistency in harassment, hate, and violence categories.
- Sexual content is the most divergent category.
- Context matters: OPs and certain country combinations show elevated disagreement.
- Practical recommendation: use calibrated thresholds, ensemble approaches, or targeted human review for categories with high disagreement.

Notes on inference and reporting

- All inferential tests reported are paired by post (paired t-test reported with degrees of freedom $= N - 1$).
- Cohen's d is reported for the paired t-test to communicate effect size.

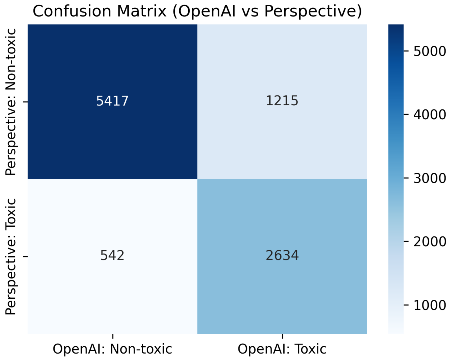

Figure 5: Heatmap of confusion matrix for OpenAI vs. Perspective classifications at threshold 0.5.

Post type	Disagreement rate (%)
Reply	17.85
OP (thread starters)	20.90

Table 4: Disagreement rates for OP posts vs. replies (binary threshold = 0.5).

- A 95% confidence interval for Pearson's r is reported using Fisher's z-transform.
- No multiple-comparisons correction (e.g., Bonferroni or BH) was applied to category-wise correlation tests; readers should interpret per-category p-values in that context (we report raw p-values).
- All the five tables and the six main figures referenced here are supplied in the Results section; the remaining 21 figures and 2 tables are provided in the Appendix.

Discussion and Implications

This study systematically compared toxicity assessments from the OpenAI Moderation API and the Google Perspective API across a large corpus of politically charged discourse from 4chan's /pol/ board. While the two systems demonstrate strong overall alignment (agreement rate $\approx 82\%$, $\rho = 0.82$), the divergences identified carry significant methodological and applied implications.

Implications for Automated Moderation

The asymmetric error structure (higher false positives for OpenAI, higher false negatives for Perspective) indicates distinct moderation philosophies: OpenAI is more sensitive to borderline or implicitly toxic content, whereas Perspec-

Country	Disagreement rate (%)
India (IN)	31.25
Costa Rica (CR)	26.67
Norway (NO)	23.33
Romania (RO)	22.41
Russia (RU)	22.02

Table 5: Top countries with highest disagreement rates between APIs.

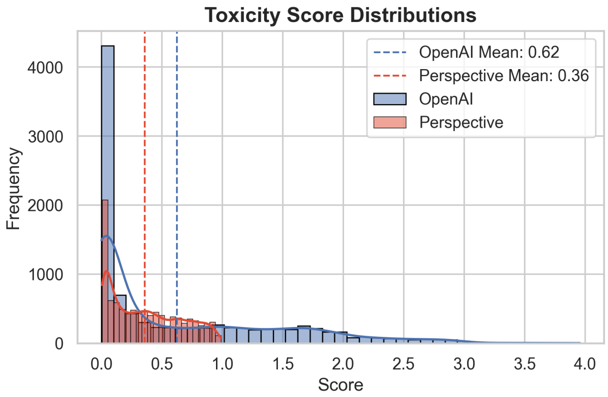

Figure 6: Overall toxicity score distributions for OpenAI and Perspective.

tive is more conservative and less likely to flag ambiguous language. In practice, these differences can translate into markedly different user experiences. A platform integrating OpenAI may suppress more content preemptively, whereas a Perspective-driven pipeline may leave more borderline cases visible. Both approaches carry trade-offs between over-moderation (risking perceived censorship) and under-moderation (risking harm from unflagged content).

The weaker cross-model alignment in the sexual content category underscores the need for calibrated or category-specific thresholds when deploying these systems operationally. Without such adjustments, one model may appear systematically "stricter" or "looser," not due to accuracy alone, but rather to divergent calibration and definitions. This is particularly consequential for platforms with legal or reputational sensitivities around sexual material.

Implications for Research

From a methodological perspective, the high correlations but significant mean differences (paired t-test: $t = 43.76$, $p < 0.0001$, Cohen's $d = 0.442$) suggest that moderation systems should not be treated as interchangeable sources of ground truth. Using one API's scores as a proxy for toxicity may bias conclusions in studies of online discourse, especially in politicized contexts. Researchers must explicitly acknowledge the calibration and design assumptions embedded in each model.

The category mapping results also highlight an opportunity: harassment, hate, and violence categories show consistently high alignment across systems, suggesting these constructs are operationalized more similarly across moderation architectures. In contrast, categories with weaker alignment (sexual content, context-dependent cases) are promising targets for deeper qualitative analysis and for building human–machine hybrid labeling pipelines.

Implications for Platform Governance and Policy

The country and thread-level divergences point to broader sociotechnical concerns. Automated moderation systems, when applied globally, implicitly encode cultural and contextual biases. Our results show that disagreement rates exceeded 20% in several national contexts — e.g., India (31.25%), Costa Rica (26.67%), Norway (23.33%), Romania (22.41%), Russia (22.02%), and Brazil (21.52%) — while Germany had the lowest disagreement rate among large-sample countries (13.00%), with only Cyprus lower overall (8.11%), underscoring the risk of misalignment between moderation algorithms and local cultural norms. For platform governance, this suggests that a single global threshold is unlikely to be adequate; regionally adaptive calibration or human-in-the-loop review may be required for fairness and legitimacy.

These findings also inform policy debates around algorithmic transparency. Demonstrating that two widely used commercial moderation APIs disagree systematically—even on the same dataset—illustrates the opacity and subjectivity of toxicity detection. Regulators and platforms alike must grapple with the fact that "toxicity" is not a fixed technical construct but a negotiated boundary that depends on model design choices.

Limitations and Future Work

Several limitations temper these findings. First, this study analyzed a single platform (/pol/), which is highly politicized and linguistically distinctive; results may differ in mainstream or less adversarial communities. Second, our analyses relied on a binary threshold of 0.5; alternate thresholds or continuous calibration may yield different agreement structures. Third, no multiple-comparisons correction was applied in category-wise tests; while correlations are strong, p-values should be interpreted conservatively. Finally, this analysis remains quantitative; qualitative exploration of disagreement cases would add important nuance.

Future work should pursue three directions: (1) extending cross-model comparisons to additional platforms and languages, (2) integrating human annotation to benchmark API outputs against socially grounded judgments, and (3) developing ensemble or threshold-adaptive moderation systems that exploit complementary strengths of multiple models.

Key Takeaway

In sum, while OpenAI and Perspective broadly agree on what constitutes toxic content, their systematic divergences in sensitivity, calibration, and context-dependence highlight the need for careful deployment, critical interpretation, and adaptive governance. Automated moderation is not a solved technical problem but an ongoing socio-technical negotiation—one where empirical comparisons, such as those presented here, can guide both research and policy toward more responsible design.

Generative AI Usage Statement

In accordance with the Yang Lab guidelines on transparency, this project made use of generative AI tools to support specific stages of the workflow.

Tools Used: Microsoft Copilot and OpenAI ChatGPT-5.

Role in the Research Process:

- *Code Assistance:* Drafted, refined, and debugged Python scripts in the `src/` directory, including:
  - `main.py` – Orchestration of the full pipeline with logging and reproducibility safeguards.
  - `data_collection.py` – 4chan JSON API integration with rate limiting, error handling, and metadata capture.
  - `processing.py` – Data cleaning, normalization, and preparation for statistical analysis.
  - `api_integration.py` – Querying the OpenAI Moderation API and Google Perspective API with structured output.
  - `analysis.py` – Statistical tests (Pearson/Spearman correlation, paired t-tests, χ^2 tests) and visualization generation.
- *Analytical Guidance:* Recommended appropriate statistical methods, advised on interpretation of results, and suggested visualization strategies.
- *Report Drafting:* Assisted in structuring and refining the *Methodology*, *Results*, and *Discussion of Implications* sections, and in improving figure/table captions for clarity and consistency.
- *Appendix Preparation:* Generated LaTeX code for the appendix to ensure consistent figure numbering (A1–A21) and cross-referencing.

Integration into the Final Report: All AI-generated code, text, and suggestions were reviewed, edited, and validated by the author to ensure technical accuracy, alignment with the dataset, and adherence to the research objectives. No outputs were used without human verification. The author retains full responsibility for the design, execution, and interpretation of the study.

Appendix: Extended Toxicity Analysis Figures

This appendix presents the complete set of supplementary figures referenced in the Results section. These include per-category score distributions. While the main text highlights representative examples, Table 6 Table 7 and the following figures (A1–A21) to provide deeper insight into calibration, sensitivity, and distributional behavior across all toxicity dimensions.

Country	Disagreement rate (%)
India (IN)	31.25
Costa Rica (CR)	26.67
Norway (NO)	23.33
Romania (RO)	22.41
Russia (RU)	22.02
Brazil (BR)	21.52
Australia (AU)	19.06
Canada (CA)	18.74
United States (US)	17.87
Netherlands (NL)	17.65
Italy (IT)	16.67
Poland (PL)	16.22
Finland (FI)	16.07
Mexico (MX)	16.05
Japan (JP)	15.79
Argentina (AR)	15.38
United Kingdom (GB)	14.29
France (FR)	13.04
Germany (DE)	13.00
Uruguay (UY)	10.34
New Zealand (NZ)	10.26
Sweden (SE)	10.26
Cyprus (CY)	8.11

Table 6: Country-level disagreement rates between APIs (binary threshold = 0.5).

Illustrative Disagreement Cases

To contextualize the quantitative patterns, Table 7 presents anonymized examples of posts where the two APIs disagreed. These cases illustrate common sources of divergence, such as sarcasm, reclaimed slurs, and coded language.

Post excerpt	**OpenAI**	**Perspective**
"That's just brilliant, you moron."	Toxic	Non-toxic
"We should send them all back."	Toxic	Non-toxic
"Nice try, sweetheart ;)"	Non-toxic	Toxic

Table 7: Sample disagreement cases (labels at threshold 0.5).

OpenAI Category Distributions

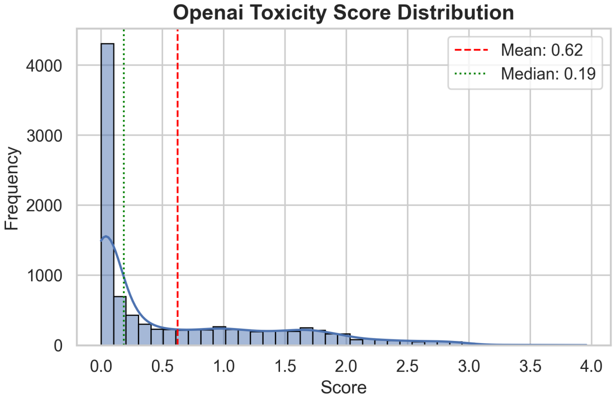

Figure A1: OpenAI toxicity score distribution.

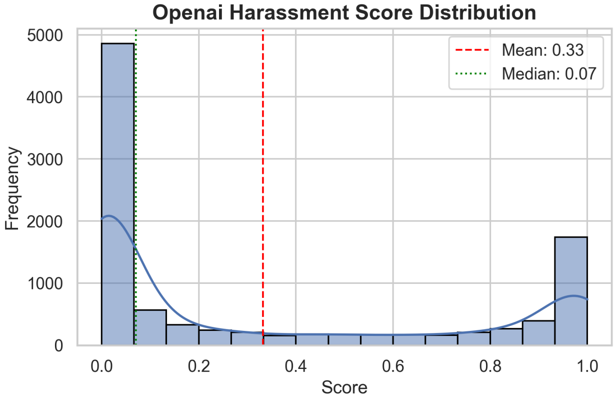

Figure A2: OpenAI harassment score distribution.

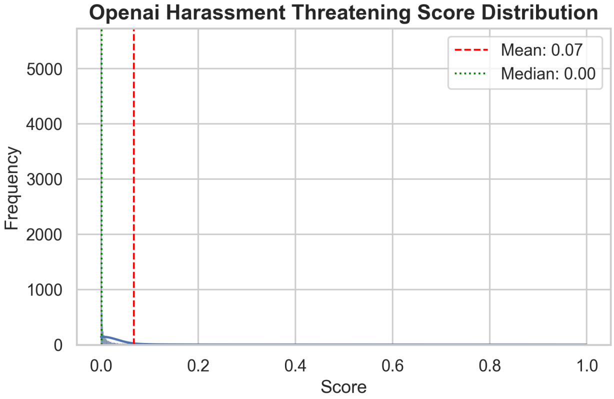

Figure A3: OpenAI harassment-threatening score distribution.

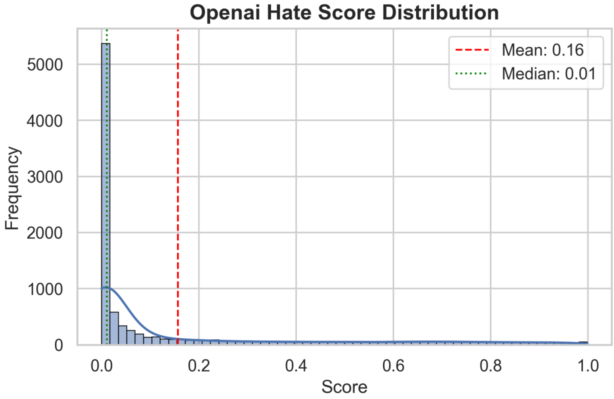

Figure A4: OpenAI hate score distribution.

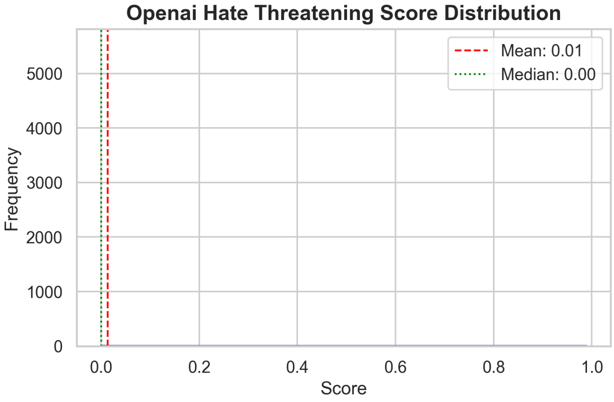

Figure A5: OpenAI hate-threatening score distribution.

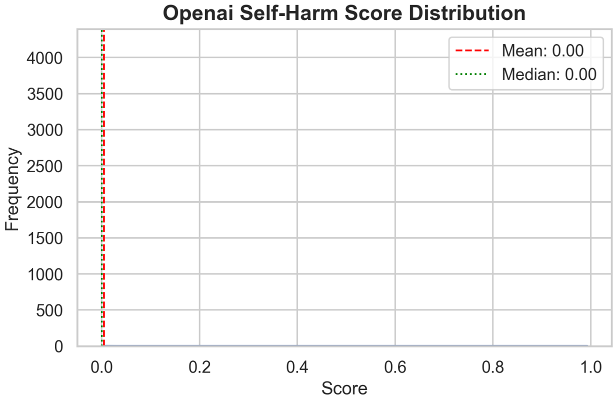

Figure A6: OpenAI self-harm score distribution.

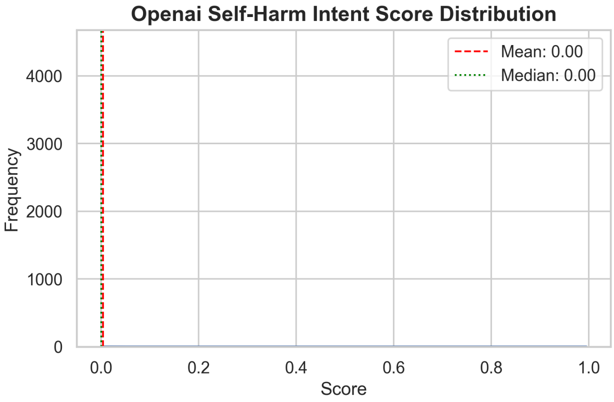

Figure A7: OpenAI self-harm-intent score distribution.

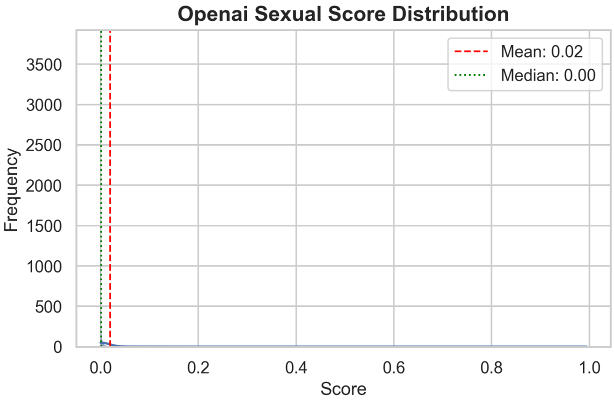

Figure A8: OpenAI sexual score distribution.

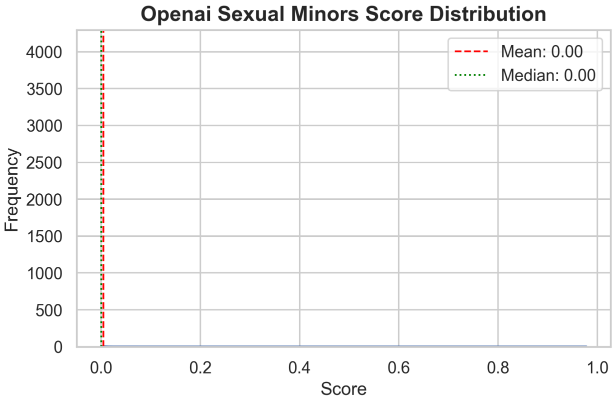

Figure A9: OpenAI sexual-minors score distribution.

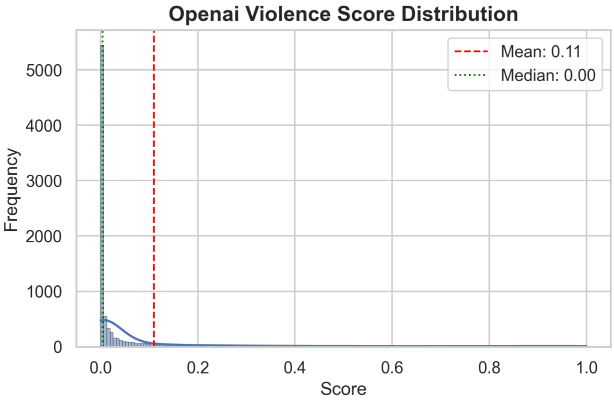

Figure A10: OpenAI violence score distribution.

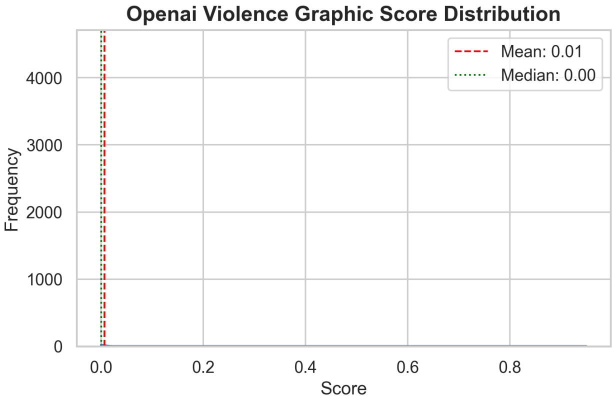

Figure A11: OpenAI violence-graphic score distribution.

Perspective API Category Distributions

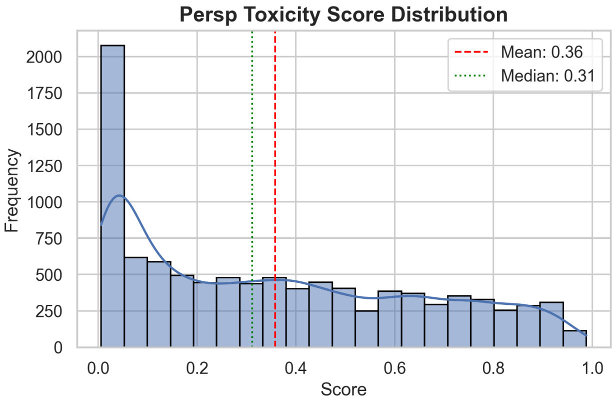

Figure A12: Perspective toxicity score distribution.

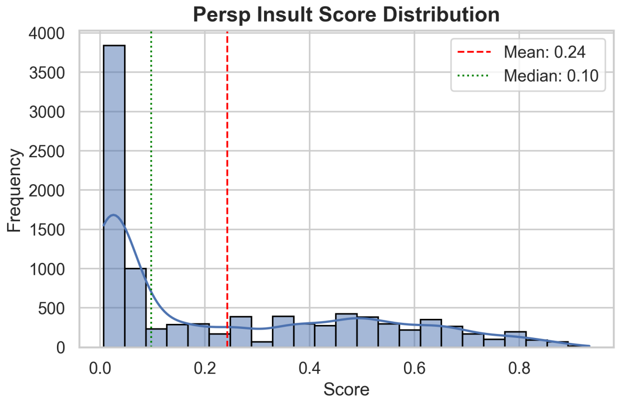

Figure A13: Perspective insult score distribution.

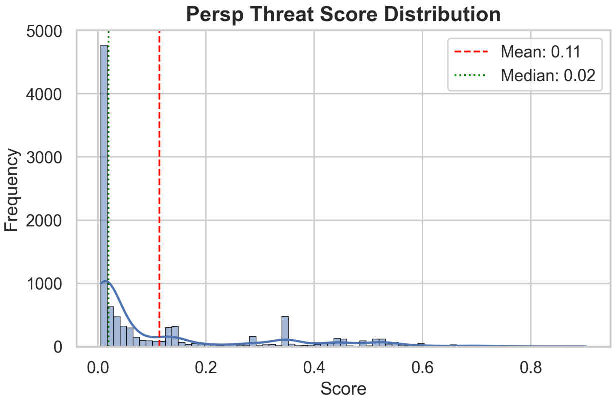

Figure A14: Perspective threat score distribution.

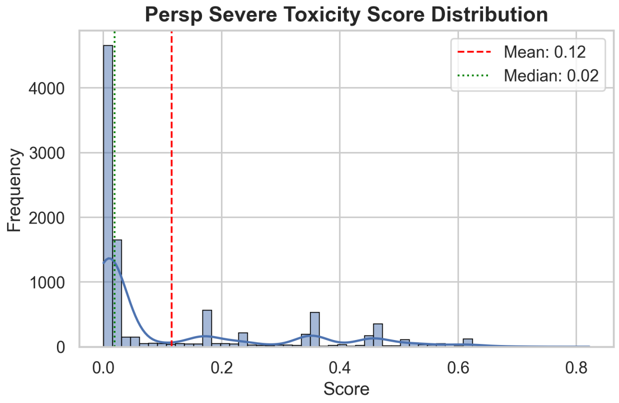

Figure A15: Perspective severe-toxicity score distribution.

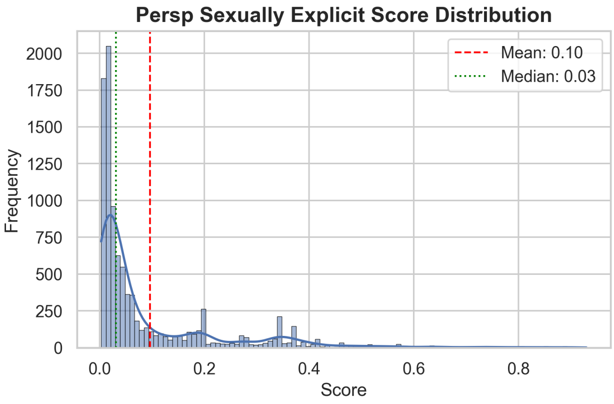

Figure A16: Perspective sexually-explicit score distribution.

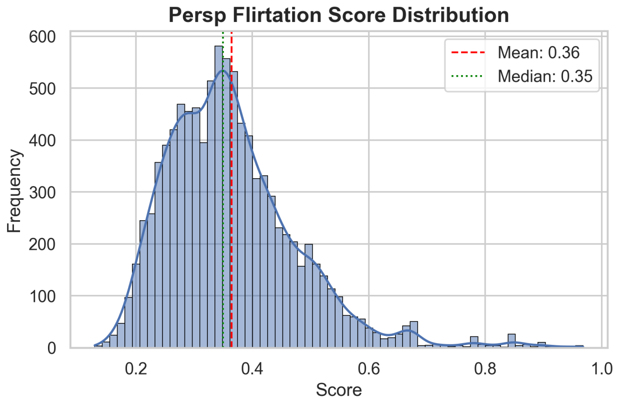

Figure A17: Perspective flirtation score distribution.

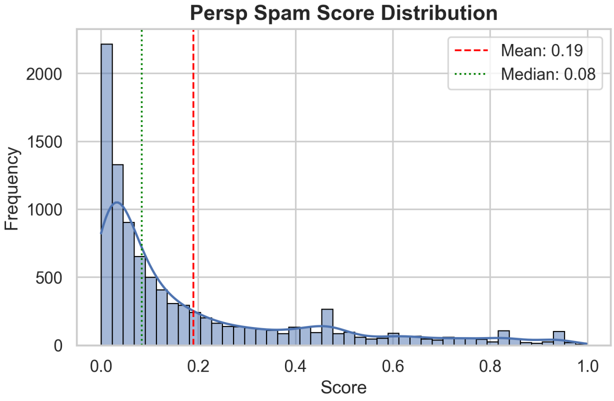

Figure A18: Perspective spam score distribution.

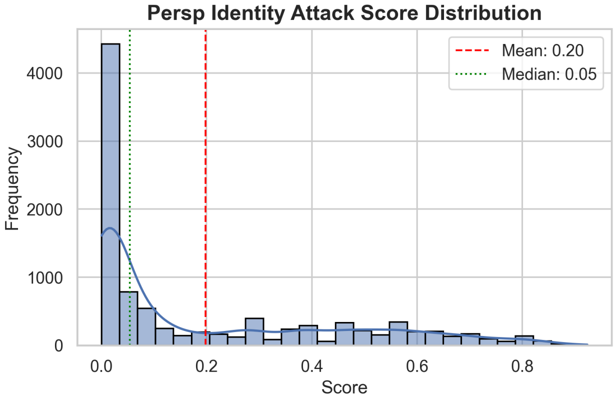

Figure A19: Perspective identity-attack score distribution.

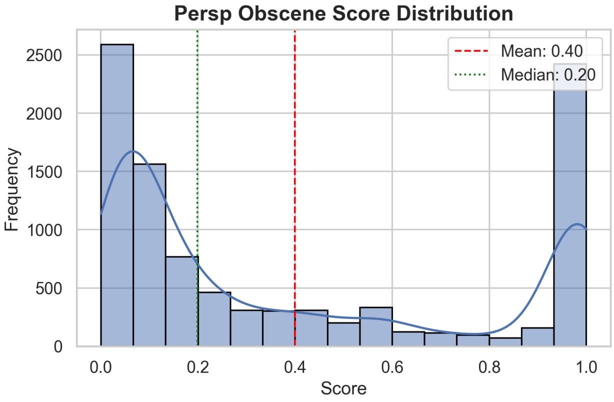

Figure A20: Perspective obscene score distribution.

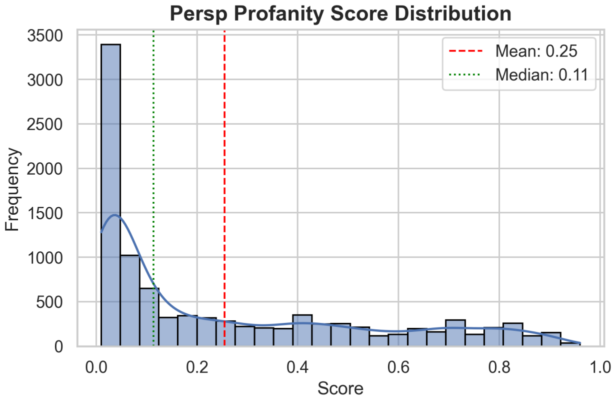

Figure A21: Perspective profanity score distribution.