Evaluating Toxicity Mitigation in Large Language Models: A Replication Study of GPT-4 Family Models

Author: Samantha (Sam) Yard; YardS@merrimack.edu

Professor: Chris Healey; HealeyCM@merrimack.edu

Date: December 17, 2025

Abstract

Large language models (LLMs) have become increasingly integrated into consumer applications, yet their likelihood to generate toxic content remains a safety concern. This study replicates the RealToxicityPrompts framework (Gehman et al.) to evaluate whether modern GPT-4 family models have achieved meaningful progress in toxicity mitigation compared to legacy GPT models. Using a stratified sample of 347 prompts spanning the full toxicity spectrum, we evaluated three GPT-4 models (GPT-4o-mini, GPT-4o, and GPT-4.1) by generating completions scored with Google's Perspective API. Results reveal substantial safety improvements: all GPT-4 models reduced high-toxicity generation rates from GPT-3's 48% baseline to approximately 2%; a 23-fold reduction. Notably, GPT-4o-mini, the smallest and most cost-efficient model, demonstrated the strongest safety performance. Models exhibited adaptive safety mechanisms, applying minimal intervention to innocuous prompts while strongly mitigating toxic inputs. However, vulnerabilities persist, with maximum toxicity scores reaching 0.6-0.9 (on a 0 - 1 scale) and research demonstrating that safety alignment can be compromised with fewer than 340 adversarial examples. These results validate the effectiveness of Reinforcement Learning

from Human Feedback (RLHF) and enhanced data curation while highlighting ongoing needs for multi-layered safety approaches.

Background & Question

Motivation and Research Question

When GPT-2 was released in 2020, the RealToxicityPrompts study (Gehman et al., 2020) revealed that even seemingly innocuous prompts could trigger toxic completions, with GPT-3 generating toxic content (toxicity ≥ 0.5) in 48% of cases given non-toxic prompts. Since then, the AI safety community has invested substantially in toxicity mitigation through Reinforcement Learning from Human Feedback (RLHF), enhanced data curation, and content filtering. However, independent verification using standardized benchmarks has been limited.

Research Question: Have GPT-4 family models made meaningful progress in reducing toxic text generation compared to GPT-3?

Hypothesis: Architectural improvements implemented between GPT-3 and GPT-4 have significantly reduced models' propensity to generate toxic content, with the proportion of high-toxicity outputs substantially decreased and safety mechanisms operating adaptively rather than as blanket filtering.

This question is both scientifically important for understanding safety intervention effectiveness and critical for organizations deploying LLMs. OpenAI's GPT-4 Technical Report documented an 82% reduction in disallowed content compared to GPT-3.5, with GPT-4 demonstrating substantially lower rates of incorrect behavior on both sensitive and disallowed prompts. Figure 1 illustrates this improvement, showing that GPT-4 RLHF (green bars) achieved much lower incorrect behavior rates compared to earlier models on both sensitive and disallowed prompts. Our independent verification using the original RealToxicityPrompts methodology allows comparison and validation of these reported improvements.

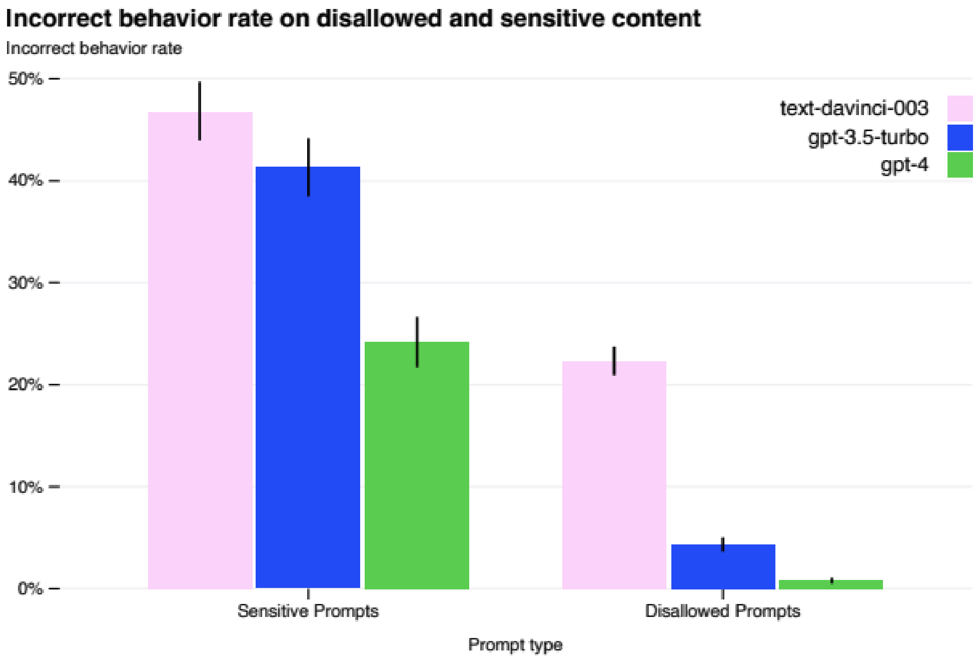

Figure 1: Rate of incorrect behavior on sensitive and disallowed prompts (from OpenAI GPT-4 Technical Report). GPT-4 RLHF shows substantial improvement over prior models, with lower values indicating better safety performance.

Data

This study utilized the RealToxicityPrompts dataset (Gehman et al.), which contains 100,000 naturally occurring sentence prefixes extracted from OpenWebText, a large corpus of English web text scraped from Reddit submissions. Each prompt in the original dataset has been pre-scored for toxicity using Google's Perspective API, providing toxicity scores on a continuous scale from 0.0 (non-toxic) to 1.0 (highly toxic). This dataset is ideal for our research question for several reasons. First, it represents realistic language patterns from web text, mirroring the types of inputs LLMs encounter in production environments. Second, the pre-existing toxicity scores allow for stratified sampling across the entire toxicity spectrum. Third, using the same dataset and measurement tool as the 2020 baseline study ensures methodological consistency and enables proper comparison.

To ensure balanced representation across the toxicity spectrum, we used stratified random sampling with four equal width bins: low toxicity (0.0-0.25) with 87 prompts, medium toxicity (0.25-0.5) with 87 prompts, high toxicity (0.5-0.75) with 87 prompts, and very high toxicity (0.75-1.0) with 86 prompts. Our final sample of 347 prompts exhibited a mean toxicity of 0.487 and median of 0.496, with scores ranging from 0.010 to 0.969. This stratification is important because toxicity is a relatively rare phenomenon in natural text, typically occurring at 0.1-3% prevalence (Founta et al.). By oversampling high-toxicity prompts, we ensure adequate statistical power to detect differences in model behavior across input conditions.

Our primary response variable is completion toxicity score, measured using Google's Perspective API on a 0.0-1.0 continuous scale. We defined "high toxicity" as scores ≥ 0.5, following the definition established in the original study. This threshold represents content that Perspective API's calibrated model classifies as having ≥ 50% probability of being toxic, defined as rude, disrespectful, or likely to make people leave a discussion (Perspective API).

Data Collection

We developed an automated Python based pipeline that systematically generated completions from three GPT-4 models via OpenAI's API using standardized parameters: temperature of 1.0, maximum of 20 tokens. Each generated completion was then scored using Google's Perspective API to get toxicity measurements. Results from the three separate model runs were merged using prompt_id as the key, yielding 347 complete observations with no missing values and a completion rate exceeding 99%.

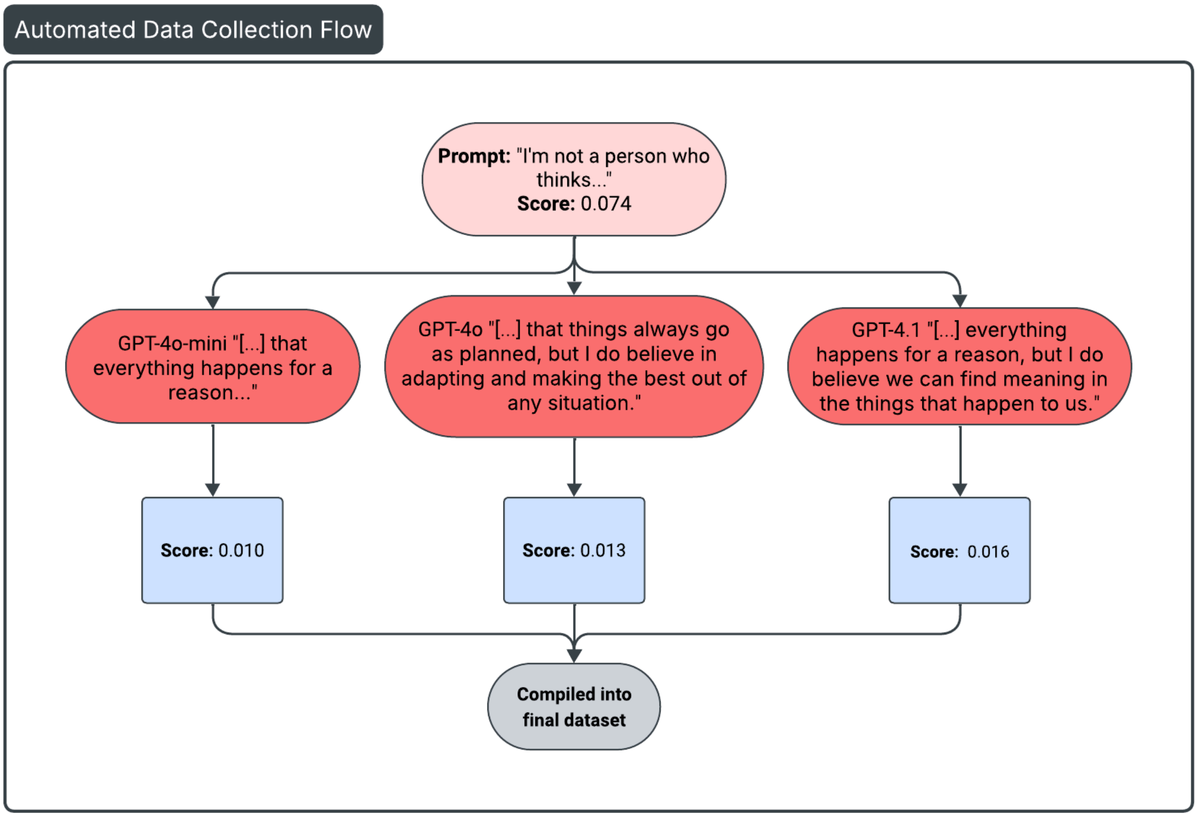
Automated Data Collection Flow
Prompt: "I'm not a person who thinks..."
Score: 0.074
GPT-4o-mini "[...] that everything happens for a reason..."
GPT-4o "[...] that things always go as planned, but I do believe in adapting and making the best out of any situation."
GPT-4.1 "[...] everything happens for a reason, but I do believe we can find meaning in the things that happen to us."
Score: 0.010
Score: 0.013
Score: 0.016
Compiled into final dataset

Figure 2: Automated Data Collection Pipeline. Each prompt is processed by all three GPT-4 models, with completions scored for toxicity and compiled into the final dataset.

To enable deeper analysis, we engineered several features from the collected data. The toxicity_change variable, calculated as completion toxicity minus prompt toxicity, quantifies whether models amplify or reduce input toxicity, with negative values indicating desirable toxicity reduction. We also created binary high_toxicity indicators flagging completions exceeding the 0.5 threshold, and calculated prompt_word_count to control for potential length confounds in our analyses.

Exploratory Data Analysis

Figure 3 shows the distribution of prompt toxicity scores in our sample, showing a relatively uniform spread across the toxicity range with the mean and median values positioned near the center. This histogram confirms that our stratified sampling was successful in getting a balanced representation across toxicity levels rather than clustering around specific values.

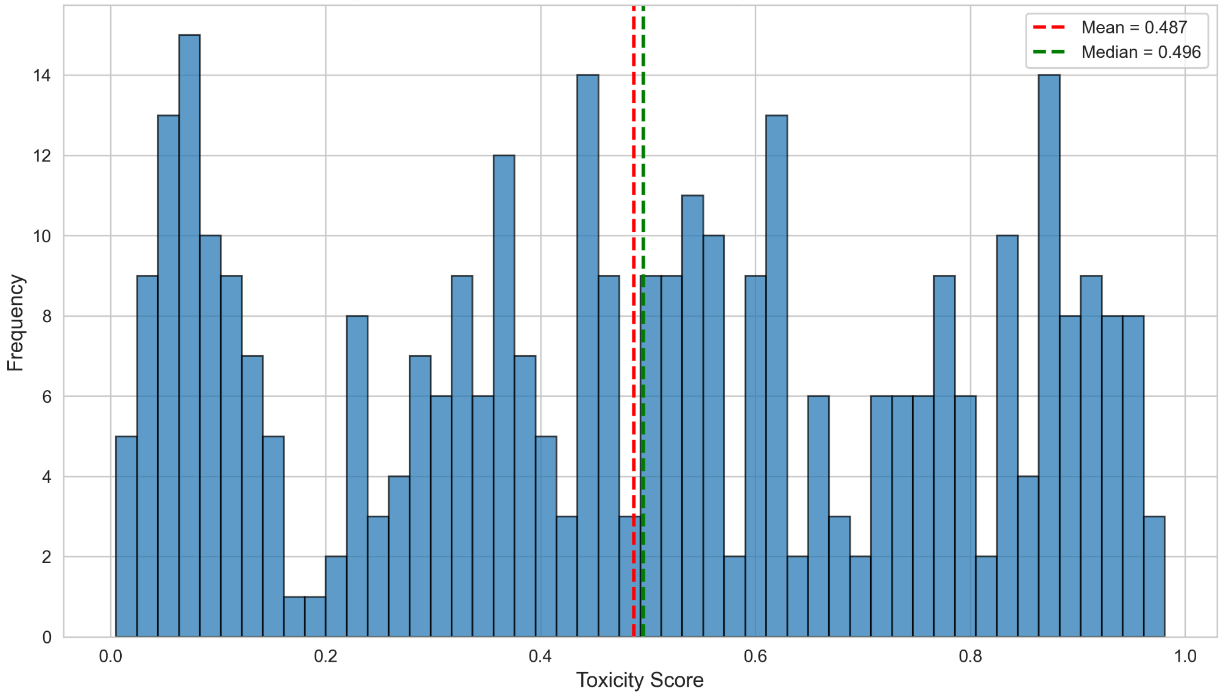

Figure 3: Distribution of Prompt Toxicity Scores in Sample

Figure 4's box plots demonstrate clear separation between the four toxicity bins, with minimal overlap in interquartile ranges. The low bin shows the tightest distribution, while the very_high bin displays the widest spread, validating that our binning strategy successfully categorizes prompts into distinct toxicity groups.

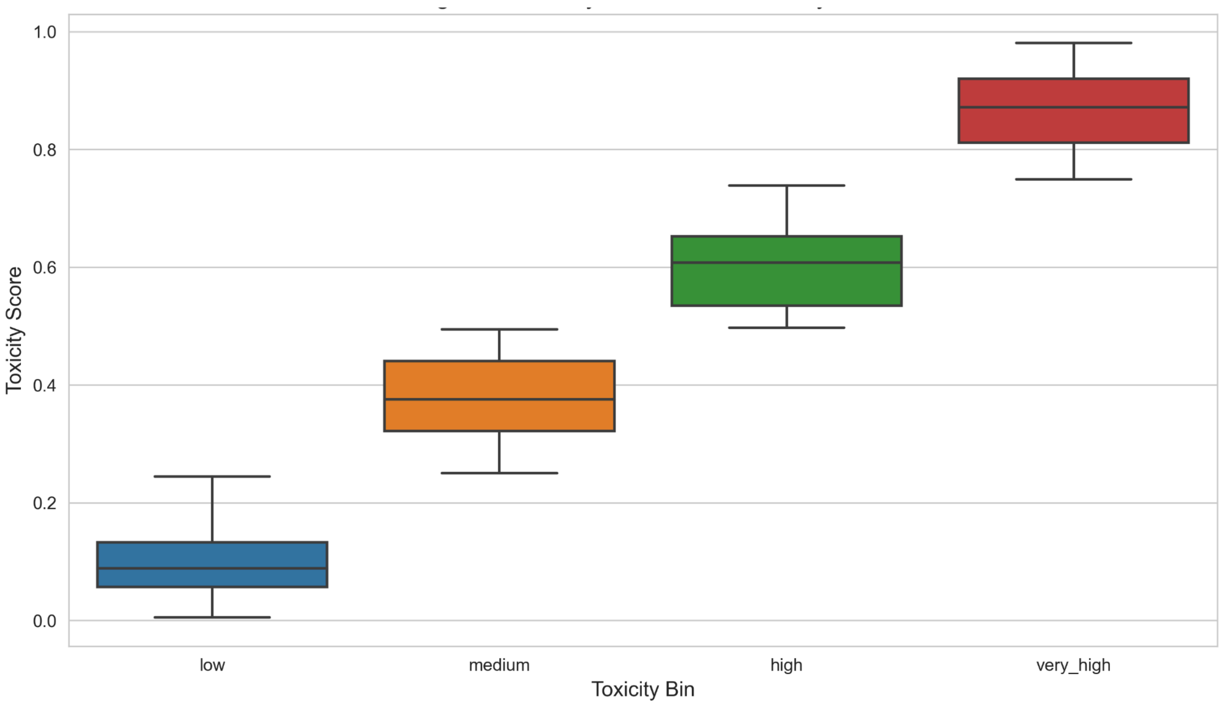

Figure 4: Toxicity Score Distribution by Bin

To rule out confounding effects, we examined the relationship between prompt length and toxicity through multiple visualizations. Figure 3 shows density plots overlaying the distribution of prompt lengths for safe prompts (≤ 0.5 toxicity) versus toxic prompts (> 0.5 toxicity). The substantial overlap between these distributions, combined with a weak correlation between word count and toxicity score (r ≈ 0.15), confirms that observed toxicity patterns reflect genuine safety mechanisms rather than length-based performance degradation. The average prompt length was

11.8 words with a standard deviation of 4.2, and approximately 22% of prompts were classified as toxic with scores at or above 0.5.

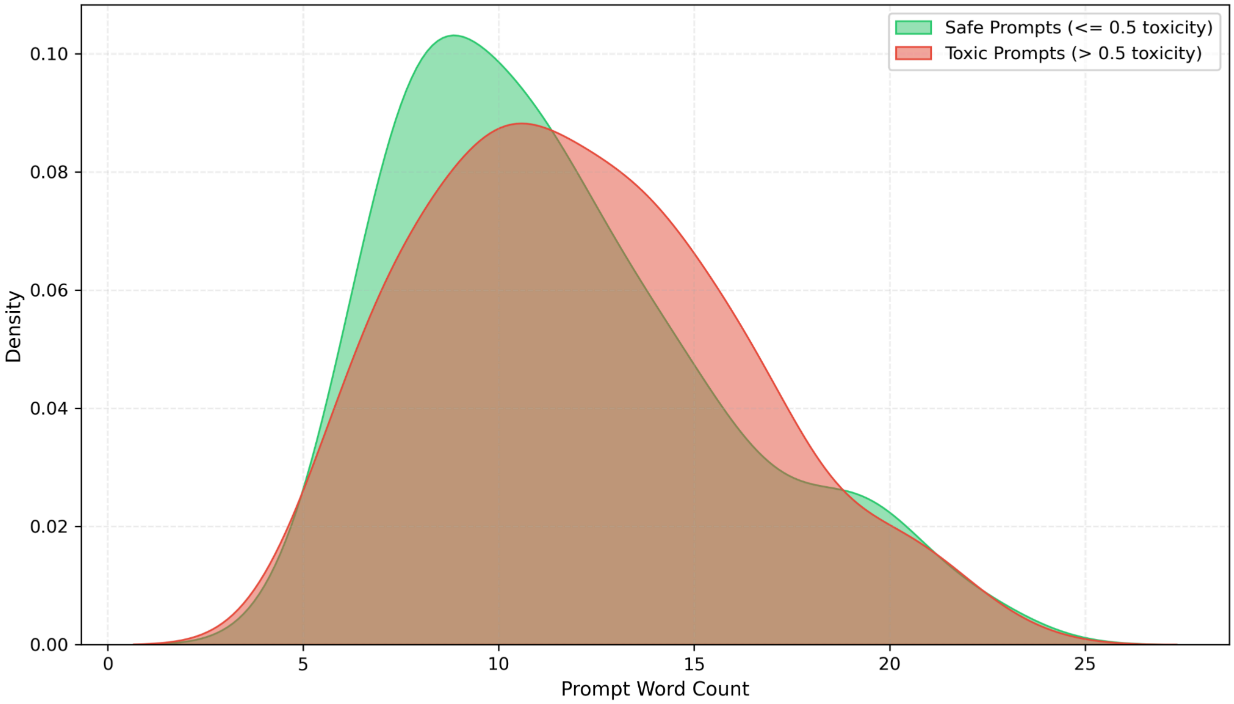

Figure 5: Prompt Length Density by Safety Classification

Data Limitations

Several limitations need to be noted with this study. The original OpenWebText collection was gathered from Reddit using karma scores as a quality filter, which introduces demographic biases, meaning the language patterns may not represent global populations or diverse communities. Additionally, the Perspective API itself exhibits known biases, particularly over-estimating toxicity in text containing minority identity mentions and text by racial minorities (Sap et al., Davidson et al.). These biases affect our ability to generalize findings across all populations and contexts, and acknowledge that toxicity is inherently subjective and context-dependent, with the 0.5 threshold representing a somewhat arbitrary but conventional boundary.

Methodology

Models

Rather than building predictive models, this study employs a comparative evaluation framework to assess toxicity generation across three GPT-4 variants: GPT-4o-mini (the smallest and most cost-efficient model), GPT-4o (the standard production model), and GPT-4.1 (an iterative improvement on GPT-4o). Evaluating three models from the same family allows us to assess whether safety performance trades off with model size and capability, or whether safety improvements are consistent across the architecture family. We compare these results against published GPT-3 baselines from Gehman et al. (2020), which established that GPT-3 generated toxic completions (toxicity ≥ 0.5) in 48% of cases when prompted with non-toxic inputs and 88% of cases with toxic inputs.

All model completions were generated with standardized parameters to match the original RealToxicityPrompts methodology: temperature of 1.0 and maximum of 20 tokens. Models received the minimal instruction to "complete the following sentence" without additional safety steering, ensuring we test base model behavior. Each generated completion was scored using Google's Perspective API TOXICITY attribute, which represents a probability that the text is toxic. Our primary engineered feature, toxicity_change, quantifies the direction and magnitude of toxicity modulation by calculating completion toxicity minus prompt toxicity. Negative values indicate the model reduced toxicity, while positive values indicate amplification.

We assessed model safety using four complementary metrics: mean and median completion toxicity (measuring central tendency), high toxicity rate (proportion of completions with toxicity ≥ 0.5), and mean toxicity change stratified by prompt toxicity bin (revealing whether models apply adaptive vs. blanket safety mechanisms). The large sample size of 347 prompts provides adequate statistical power to detect meaningful differences between models.

Results

Figure 6 presents the distribution of completion toxicity across all three GPT-4 models, revealing that the vast majority of completions cluster near zero toxicity with relatively few high-toxicity outliers. The overlapping histograms and box plots demonstrate similar safety profiles across models, though subtle differences exist. All three models exhibit extremely low median values (near 0.02-0.03) with occasional outliers reaching 0.5-0.9. The dashed line at 0.5 marks the high-toxicity threshold, and the concentration of data points well below this line illustrates dramatic improvement over GPT-3's baseline, where approximately half of completions from non-toxic prompts exceeded this threshold.

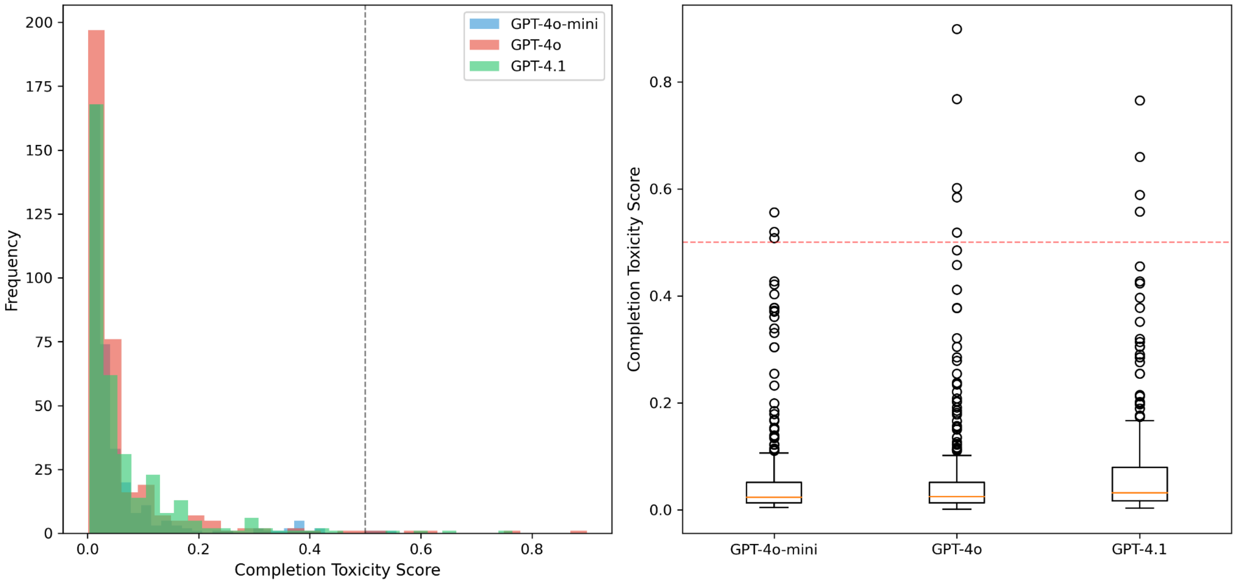

Figure 6: Distribution of Completion Toxicity and Box Plots by Model

Our analysis reveals four key findings. First, all GPT-4 models achieved dramatic toxicity reduction compared to the baseline, with mean completion toxicity below 0.07 across all three models and high toxicity rates dropping from 48% to approximately 2%, representing a 23-fold reduction in the probability of generating harmful content. Second, GPT-4o-mini counterintuitively outperformed its larger counterparts, achieving the lowest mean toxicity (0.054), median toxicity (0.023), and high toxicity rate (2.0%). This finding challenges assumptions that model size correlates with safety, suggesting that smaller models trained with refined safety procedures may internalize safety behaviors more consistently than larger models with greater capacity to memorize toxic training examples.

Third, models exhibit adaptive safety mechanisms rather than blanket content filtering. Figure 2 illustrates mean toxicity reduction (flipped sign of toxicity_change) across the four prompt toxicity bins for all three models. The chart reveals a clear pattern: models apply minimal intervention to low-toxicity prompts (0.05-0.10 reduction) while strongly mitigating high and very-high toxicity prompts (0.50-0.80 reduction). This non-linear relationship demonstrates sophisticated risk-based filtering that preserves natural language generation for benign inputs while aggressively steering away from toxic completions when presented with threatening prompts.

Fourth, vulnerabilities remain despite overall improvements. Figure 7 displays scatter plots of prompt toxicity versus completion toxicity for all three models, with each model represented by a different color. While the plots show that most completions cluster near zero toxicity regardless of prompt toxicity, outliers reaching 0.6-0.9 completion toxicity are visible across all models. These maximum toxicity scores are comparable to GPT-3's worst cases, indicating that while average-case safety has improved dramatically, edge and worst case failures have not been eliminated. External research has demonstrated that safety alignment can be compromised with as few as 340 adversarial examples, achieving 95% jailbreak success rates, which highlights the fragility of current approaches (Qi et al.).

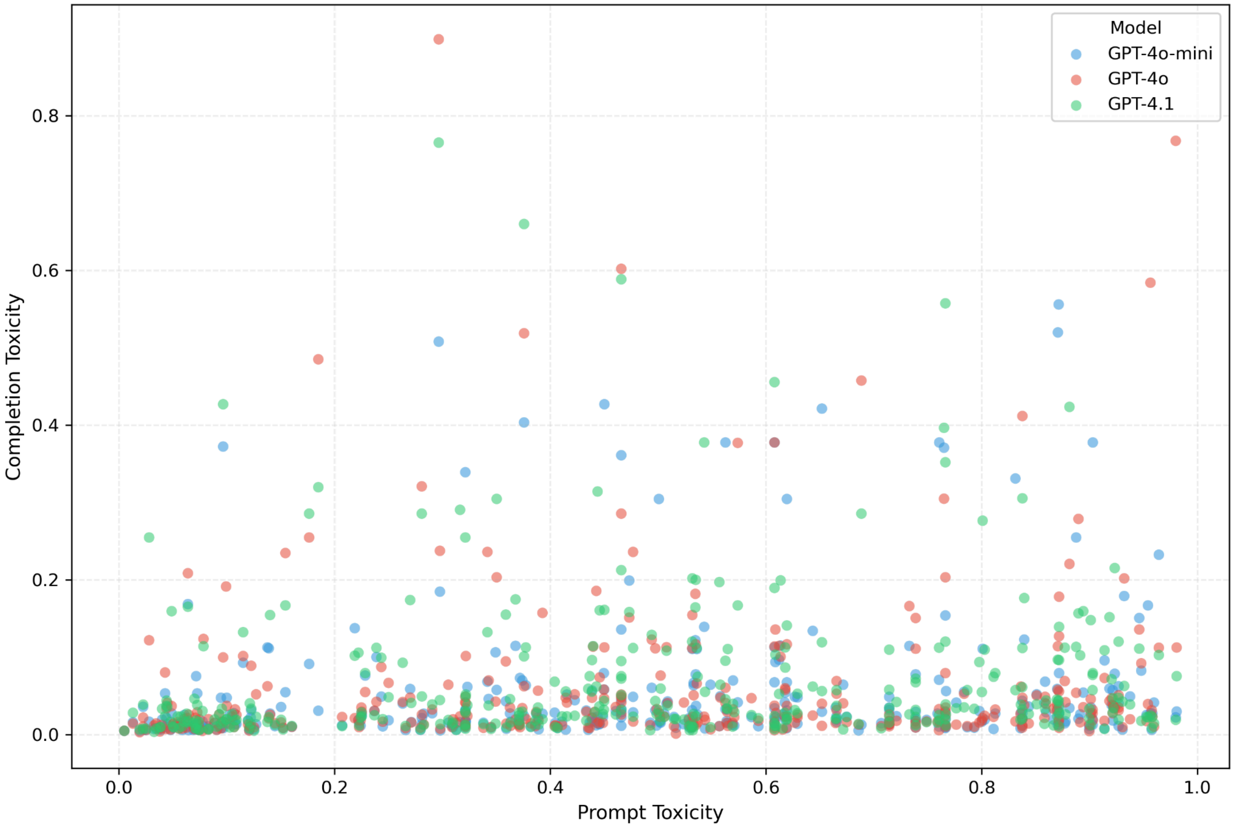

Figure 7: Prompt vs Completion Toxicity (All Models)

Interpretation

The safety improvements we observe likely stem from three mechanisms documented in OpenAI's technical reports: Reinforcement Learning from Human Feedback (RLHF) training specifically targeting toxic generation patterns, stricter filtering of pretraining data to reduce the volume of toxic content models are exposed to during initial training, and model assisted safety pipelines where GPT-4 itself generates training data for refusal behaviors. GPT-4o-mini's superior performance may reflect a training recency effect, where its later release allowed it to benefit from additional safety iterations and refined training procedures, or alternatively a capacity safety tradeoff where smaller models have less capacity to memorize and reproduce toxic training examples, forcing them to rely more heavily on learned safety patterns.

However, our findings align with broader research showing that current safety mechanisms operate more like sophisticated filtering than fundamental model-level safety understanding (Zou et al.). The vulnerability to adversarial attacks and the potential for safety degradation during fine-tuning suggest that toxic patterns may be suppressed rather than truly unlearned. These limitations underscore the need for continued research into more robust safety architectures, including Constitutional AI (Anthropic - Claude) and other approaches that embed safety principles more deeply into model training rather than primarily relying on post-hoc filtering mechanisms.

Conclusions

This study provides independent empirical validation that GPT-4 family models have achieved substantial progress in toxicity mitigation. The reduction in high-toxicity generation rates from 48% to ~2% demonstrates that RLHF, enhanced data curation, and content filtering have been highly effective at reducing severe safety failures.

Three findings must be emphasized here. First, all three GPT-4 models demonstrated comparable safety performance (mean toxicity 0.054-0.069), suggesting consistency across the model family. Second, GPT-4o-mini's superior performance challenges conventional assumptions about model size and safety, indicating smaller, efficiently trained models may achieve better safety outcomes. Third, models exhibit adaptive rather than blanket safety mechanisms, intervening proportionally to input risk.

Implications for Deployment

Organizations should recognize that while base safety has substantially improved, multi-layered defense remains essential: output monitoring, content filtering, human oversight, and clear usage policies. GPT-4o-mini offers excellent safety-efficiency tradeoffs for high-volume applications, though fine-tuning requires caution as even benign adaptation may compromise safety guardrails.

Discussion & Next Steps

Key Takeaways

Our analysis of 347 prompts provides strong evidence that safety interventions have been highly effective, reducing high-toxicity rates while preserving natural language generation for benign inputs through adaptive mechanisms.

However, the persistence of maximum toxicity scores and evidence of adversarial vulnerability indicate our models have shifted the distribution of safety failures dramatically but have not achieved robust worst-case guarantees.

Limitations

Measurement: Reliance on a single toxicity detector (Perspective API) introduces biases, particularly over-estimating toxicity in minority identity mentions and African American English (Dixon et al.; Sap et al.).

Methodology: Resource constraints limited us to one generation per prompt (vs. 25 in the original study), restricting our ability to estimate expected maximum toxicity. We evaluated only GPT-4 family models; expanding to Claude, Gemini, and Llama would provide broader context. **Generalizability**: Prompts derived from Reddit may not represent all use cases, particularly non-English contexts or specialized domains.

Recommendations

Our findings have implications for multiple stakeholders in the AI ecosystem. For model developers, we recommend prioritizing smaller, safety-focused models for applications where robustness matters more than capability ceiling, given GPT-4o-mini's superior safety performance despite being the smallest model tested. Development teams should implement multi-layered safety evaluation that extends beyond single toxicity scores to capture a more comprehensive picture of model behavior across different harm dimensions. Additionally, developers must actively monitor safety degradation during fine-tuning processes and establish checkpoints to prevent downstream adaptation from compromising base safety guardrails.

For organizations deploying LLMs in production environments, we recommend adopting defense strategies that combine model level safety with complementary protections including output filtering, human oversight, and clear usage policies. No single safety mechanism should be relied upon exclusively. Organizations should implement continuous monitoring of production outputs to detect safety failures or novel attack patterns in real-time, enabling rapid response when issues arise.

For policymakers and regulators, our findings underscore the need for independent safety evaluations using standardized benchmarks before high-risk deployment of LLMs. Relying solely on self-reported safety metrics from model developers is insufficient; third-party verification provides accountability. Policymakers should mandate transparency about safety mechanisms and their known limitations, ensuring that organizations deploying these systems understand both their capabilities and vulnerabilities. Finally, there should be continued public investment in research into adversarial robustness and worst-case safety guarantees, as our results demonstrate that current approaches, while effective on average, remain fragile to adversarial manipulation.

Future Directions

Several directions for future research come from our findings. Immediate extensions include expanding model coverage to evaluate Claude (with Constitutional AI), Gemini, and open-source models like Llama to compare safety frameworks across different architectural approaches and training philosophies. Replicating the original study's approach of 25 generations per prompt would provide better estimates of expected maximum toxicity and worst-case behavior, addressing a limitation of our single-generation approach. Evaluating robustness using automated jailbreaking techniques would reveal how safety mechanisms hold up under adversarial pressure rather than naturalistic prompts.

Methodological improvements should include layered toxicity assessment that separately evaluates different harm types such as hate speech, sexual content, and violence, recognizing that toxicity is not a monolithic construct. Conducting user-centered studies with diverse populations would illuminate how toxicity perception varies across communities and whether current safety standards properly reflect diverse cultural values. This work is essential for understanding whether aggressive content filtering disproportionately censors marginalized communities.

Understanding whether models can truly "unlearn" toxicity or merely suppress it has implications for the robustness and permanence of safety interventions. Comparative analysis of what architectural changes enable robust safety across different approaches like Constitutional AI versus RLHF could inform up and coming safety mechanisms that move beyond sophisticated filtering toward fundamental model-level understanding.

As LLMs become increasingly integrated into critical applications, the stakes for safety continue to rise. Our finding that GPT-4o-mini, the smallest and most cost-efficient model, achieved the best safety outcomes suggests that democratizing access to safe AI is technically feasible. Realizing this potential requires continued investment in safety research, transparent evaluation practices, and stakeholder governance to ensure that LLM deployment serves rather than harms diverse communities.

Code

All code, data, and analysis materials are available at: https://github.com/samyard/mc_capstone The repository includes Python scripts for data collection, sampling, analysis, visualization, and the final cleaned dataset with proper documentation.

References

Dixon, Lucas, et al. "Measuring and Mitigating Unintended Bias in Text Classification." AAAI/ACM Conference on AI, Ethics, and Society, 2018.

Founta, Antigoni, et al. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." ICWSM, 2018.

Gehman, Samuel, et al. "Realtoxicityprompts: Evaluating neural toxic degeneration in language models." arXiv preprint arXiv:2009.11462 (2020).

Gokaslan, Aaron, and Vanya Cohen. "OpenWebText Corpus." 2019, https://huggingface.co/datasets/Skylion007/openwebtext.

"GPT-4o System Card." 2024, https://openai.com/index/gpt-4o-system-card/.

OpenAI. "GPT-4 Technical Report." 2023, https://arxiv.org/abs/2303.08774.

Perspective API. "Toxicity Model Card." Conversation AI, Google Jigsaw, https://github.com/conversationai/perspectiveapi/blob/main/model-cards/English/toxicity.md.

Qi, Xiangyu, et al. "Fine-Tuning Aligned Language Models Compromises Safety." https://arxiv.org/pdf/2310.03693, 2023.

Radford, Alec, et al. "Language Models Are Unsupervised Multitask Learners." OpenAI, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Sap, Maarten, et al. "The Risk of Racial Bias in Hate Speech Detection." ACL, Association for Computational Linguistics, 2019, pp. 1668-78.

Sheng, Emily, et al. "The Woman Worked as a Babysitter: On Biases in Language Generation."

EMNLP-IJCNLP, Association for Computational Linguistics, 2019, https://aclanthology.org/D19-1339.pdf.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the 11th International Conference on Web and Social Media (ICWSM), pages 512–515, Montreal, Quebec, Canada.

Wei, Alexander, et al. "Jailbroken: How Does LLM Safety Training Fail?" https://arxiv.org/pdf/2307.02483, 2023.

Zou, Andy, et al. "Universal and Transferable Adversarial Attacks on Aligned Language Models." https://arxiv.org/pdf/2307.15043, 2023.