[Re] Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Web-Scale Datasets for Responsible LLMs

Anonymous authors
Paper under double-blind review

Abstract

Web-scale pretraining datasets for large language models (LLMs) contain harmful content, and existing binary filtering approaches either fail to prevent toxicity leakage or remove high-quality contextual material, reducing downstream utility. To address this trade-off, Mendu et al. (2025) introduced a three-dimensional safety taxonomy and the Topical and Toxic Prompt (TTP). However, their evaluation relied primarily on proprietary models, limiting reproducibility and scalability. In this work, we address this limitation by reproducing and evaluating the proposed taxonomy and prompt-based framework using several open-weight models. We benchmark TTP across a comprehensive cohort of open-weight architectures, identifying Llama 3.3 70B as a cost-effective alternative that matches GPT-4o. We replicate the original web-scale harm analysis on Common Crawl, C4, and FineWeb, reproduce the HAVOC benchmark to measure toxicity leakage, and extend the study by deploying HarmFormer at the scale of 1 million pages per dataset. We also report computational and environmental costs for each pipeline. Our results confirm that open-weight models are viable for transparent, reproducible dataset curation. However, we observe that the Safe-Topical boundary depends on the model chosen. Specifically, open-weight models consistently detect at rates twice as high as originally reported, particularly in Ideological Harm.

1 Introduction

Large Language Models (LLMs) are pretrained on massive web-scale datasets such as Common Crawl (Common Crawl Foundation, 2024), C4 (Raffel et al., 2020) and FineWeb (Penedo et al., 2024). While this enables broad generalization, indiscriminate data collection introduces harmful content, motivating the use of safety filters throughout the model life-cycle (Dong et al., 2025). Manual dataset curation at this scale is infeasible, and overly aggressive filtering risks removing high-quality educational and journalistic material (Nguyen et al., 2025).

Mendu et al. (2025) proposed a novel three-dimensional safety taxonomy that distinguishes between *Safe*, *Topical*, and *Toxic* content across five harm categories: *Hate & Violence, Ideological Harm, Sexual Content, Illegal Activities and Self-Inflicted Harm*, enabling preservation of contextual discussions upon sensitive topics.

Their Topical and Toxic Prompt (TTP) outperformed binary classifiers on multiple benchmarks but relied primarily on proprietary models, limiting reproducibility and raising barriers to adoption by the open research community. This paper reproduces and evaluates the TTP framework under open and reproducible conditions. We benchmark the prompt across a diverse set of open-weight models, including Llama, Mistral, and Gemma, while assessing its robustness on standard moderation datasets and replicating large-scale dataset analyses using publicly available corpora. We further extend this analysis by deploying HarmFormer, the fine-tuned LongFormer classifier proposed by Mendu et al. (2025), to classify harmful content across 3 million webpages per dataset, enabling a statistically robust comparison against both the original findings and our TTP-based results. Additionally, we reproduce the HAVOC benchmark introduced in the original

work to measure the leakage rate of open-source models across different harm categories and to characterize the most prevalent harms in openly distributed datasets. To promote transparency and reproducibility, we report the computational and environmental costs associated with these experiments, enabling a practical assessment of the feasibility of deploying such safety pipelines at scale.

2 Scope of reproducibility

Ensuring that web-scale pretraining datasets are free of harmful content is a critical and largely unsolved challenge in responsible LLM development. In the context of Algorithmic Accountability (Raji et al., 2020; Wieringa, 2020), developers (actors) have an obligation to justify the conduct of their systems to the broader society (forum). This necessitates robust Risk Management frameworks and Empirical Performance evaluations techniques essential for guaranteeing that a system behaves "as promised", regarding safety and fairness objectives. AI safety guardrails, designed to detect and filter harmful or toxic content, serve as a critical architectural layer to bridge the "accountability gap" between a designer's control and the algorithm's complex behavior. Mendu et al. (2025) proposed four primary contributions that address these challenges:

A new Three Dimensional Taxonomy:

A three-dimensional taxonomy is proposed to detect harmful content across the five harm categories. This framework addresses the limitations of traditional binary classification by introducing a new *Topical* dimension, defined as "contextual discussions requiring preservation". They claim that:

- The three-dimensional taxonomy outperforms traditional binary approaches in detecting harmful content.

Moderation Tools:

They developed TTP-eval, an "annotated benchmark for long-text generation"; TTP, a prompt developed to follow the new Taxonomy on GPT-4o; and a LongFormer model (Beltagy et al., 2020), called HarmFormer, fine-tuned by Mendu et al. (2025) on a subset of the 3 Million web pages dataset discussed in Section 4.3, which filters harmful content according to the new Taxonomy. They compared these models on the OpenAI moderation dataset (Markov et al., 2023) along with other models like: PerspectiveAPI (Lees et al., 2022) and LlamaGuard (Inan et al., 2023). The key findings reported by Mendu et al. (2025) are as follows:

- TTP achieves substantially higher performance when paired with GPT-4 Omni compared to other models.
- TTP and HarmFormer outperform PerspectiveAPI and LlamaGuard on the OpenAI Moderation dataset.

Web-Scale Harm Analysis:

Mendu et al. (2025) systematically evaluated pretraining datasets according to the new Taxonomy using TTP, aiming to quantify harmful content at scale. As result of the analysis, it is stated that:

- Large-scale datasets commonly used for pretraining LLMs contain a measurable fraction of toxic content with Sexual harm being the most prevalent category.

HAVOC Benchmark:

The HAVOC benchmark proposes an evaluation framework for measuring toxic content generation in language models across prompts that vary both in severity and harm category. In the original study, Mendu et al. (2025) assessed a range of open-source LLMs and reported two key results:

- Across their prompt suite, open-source models produced toxic outputs for 76% of prompts, a substantially higher rate than what is typically reported in prior toxicity benchmarks.

- Among the harm categories considered, sexual harm has been found to be the most frequent source of toxic generations.

Since Mendu et al. (2025) did not provide the code for reproducibility scopes, we implemented the framework from scratch to reproduce their work across three key pillars: analyze harmful content in pretraining datasets, evaluate moderation tools on standard benchmarks and measure leakage rates via the HAVOC benchmark. We further expanded this scope to test open-source models, achieving results comparable to GPT-4o.

Code availability. An anonymized implementation of all experiments is available at `https://anonymous.4open.science/r/safer-pretraining-reproduction/`.

3 Datasets

3.1 TTP-Eval Dataset

The TTP-Eval dataset was constructed by sampling one million random web pages from Common Crawl and applying a high-recall filter to identify potentially harmful content. This resulted in a filtered corpus of approximately 50,000 candidate documents. From this set of document 491 webpages were reportedly selected via stratified sampling across harm categories for manual annotation by two domain experts with over five years of experience in harm detection, achieving a Krippendorff's Alpha of 0.77 indicating substantial inter-annotator agreement Mendu et al. (2025). We utilized the publicly available dataset[1], which, however, contains only 393 annotated webpages rather than the reported 491, where only 94 records (23.9%) exhibit an Intent label in at least one category and are thus classified as toxic.

3.2 Common Crawl (CC), C4 and FineWeb

Our analysis utilizes the three datasets relied upon by the original paper: **Common Crawl (CC)** (Common Crawl Foundation, 2024), **C4** (Raffel et al., 2020), and **FineWeb** (Penedo et al., 2024). We access C4 and FineWeb via Hugging Face (Lhoest et al., 2021), while CC is obtained directly from the `CC-MAIN-2024-46` WARC archives, filtering by language[2]. Because the original paper did not provide detailed information on the original sampling procedure, we extracted two random samples per dataset: a 10,000-record sample used for TTP-based analysis, and a 1,000,000-record sample used for HarmFormer-based analysis. A length-based analysis of our samples revealed high variance, with outliers exceeding 40,000 characters. To satisfy our 16k token context window constraint, we implemented a filter to retain only records between 200 and 40,000 characters. Table 7 summarizes dataset statistics before and after our preprocessing. These datasets have also been used to create the **Havoc Benchmark Dataset**, comprised of 10,376 samples across the dimensions and harms identified by the TTP.

3.3 OpenAI Moderation Dataset

The dataset created by Markov et al. (2023) is an established industry baseline for content policy compliance (Inan et al., 2023; Ghosh et al., 2024; Franco et al., 2025). The evaluation set consists of 1,680 samples, divided into 1,158 safe and 522 toxic entries. These samples are labeled according to a hierarchical taxonomy comprising the following categories: *Sexual*, *Hate*, *Violence*, *Harassment*, *Self-Harm*, *Sexual/Minors*, *Hate/Threatening* and *Violence/Graphic*.

4 Methods

4.1 Topical and Toxic Prompt (TTP)

The Topical and Toxic Prompt (TTP) is a prompt designed to classify text according to the three dimensional taxonomy (Safe, Topical, Toxic) across the five harm categories. The original paper developed TTP through

[1] `https://github.com/themendu/TowardsSaferPretraining`

[2] The filtering process relied on the HTML global lang attribute, restricted to documents with the value set to English.

an iterative refinement process on a held-out development set, incorporating few-shot prompting (Brown et al., 2020) and chain-of-thought reasoning (Wei et al., 2022).

The prompt provides clear guidelines to distinguish between content that merely discusses harmful topics (Topical) and content that actively promotes or contains harmful material (Toxic). A key challenge identified by Mendu et al. (2025) was handling "reporting web pages", pages that describe harmful incidents without endorsing them, particularly in the Hate & Violence category.

4.1.1 TTP Model Selection and Evaluation

Mendu et al. (2025) employed OpenAI's GPT-4-Omni (GPT-4o) for all TTP-based experiments, including evaluation on TTP-Eval, large-scale dataset analysis and the HAVOC benchmark. Table 4 in the original paper reports TTP results on alternative models, DeepSeek-R1, Gemini 2 Flash and Gemma 2, showing substantially lower F1-scores than GPT-4o. However, given the prohibitive cost of running GPT-4o at scale, we conducted an extensive evaluation to identify open-source alternatives. Our primary objective was two-fold: first, to validate whether the TTP prompt sensibility could successfully transfer to open-weight models, thereby demonstrating the generalizability of the proposed taxonomy beyond proprietary systems, and secondly, to identify the best-performing open-source model that could fit within our computational constraints (Section 6.2) while matching GPT-4o's performance on TTP-Eval.

We replicate Mendu et al. (2025) comparative evaluation and extended it with an additional set of 11 open-weight models, spanning different architectures and parameter scales:

- **Open-Source GPT Models**: GPT-OSS 20B and GPT-OSS 120B
- **Meta Llama Models**: Llama 3.3 70B in full precision and quantized variants
- **Mistral AI Models**: Mistral Large 123B and Mixtral 8x7B
- **Google Gemma Models**: Gemma 2 27B and Gemma 3 27B
- **Qwen Models**: Qwen2.5 72B and Qwen3 32B
- **DeepSeek Models**: DeepSeek-R1 70B and DeepSeek-R1 32B

While our focus is on open-weight models, GPT-4o and Gemini 2 Flash are tested as reference baselines.

Baseline: Perspective API We evaluate Google's Perspective API (Lees et al., 2022) on TTP-Eval to compare the TTP approach against established commercial content moderation systems. Following the original paper's methodology, we address Perspective API's limitation in handling long-form text by chunking each webpage into segments of 500 characters and taking the maximum toxicity score across all chunks as the final prediction. We applied a threshold of 0.4 on the API's scalar output (ranging from 0 to 1) to classify content as toxic, using the same threshold Mendu et al. (2025) determined from the development set. Moreover, since the TTP-eval dataset has for each page a set of languages, we adapted the format of the language, excluding the one not recognized by the API, and included them in the request, thus avoiding the prone-to-error language inference system of the model.

Inference Configuration For all models, we use greedy decoding as the default configuration. We also conduct additional experiments at temperature=0.8 for selected models (GPT-4o and Llama 3.3) to assess potential performance gains from sampling-based decoding.

Open-source models are deployed locally (see Section 6.2) via the Ollama framework with a 16k-token context window, sufficient for the longest TTP-Eval webpages. For Gemma 2 and Gemma 3, the context is limited to 8k tokens. GPT-4o and Gemini 2 Flash are accessed through the OpenRouter API using the OpenAI client with default settings.

The TTP prompt from the original GitHub repository is used as-is, preserving instructions and examples. Since the released prompt contained OpenAI-specific formatting tokens, it has been adapted for cross-model

compatibility using LangChain, converting it into a conversational message format for few-shot evaluation while retaining semantic content. This adaptation enables multi-model evaluation but may introduce minor deviations from the original implementation.

Evaluation Protocol Following the original paper's methodology, we compute precision, recall, and F1-score specifically for the Toxic dimension across all five harm categories. For each webpage in TTP-Eval, we aggregate the model's predictions; a webpage is classified as Toxic if the model identified toxic content in any of the harm categories. This binary classification (Toxic vs. Non-Toxic) is then compared against the ground truth annotations to compute the evaluation metrics.

The best-performing model from this evaluation is selected for all subsequent experiments, including the OpenAI Moderation Dataset comparison (Section 4.2), the web-scale dataset analysis (Section 4.3) and the HAVOC experiment (Section 4.4), allowing cost-effective reproduction of the original paper's large-scale evaluations.

4.2 AI Safety Guardrails Model comparison on the OpenAI Moderation Dataset

Following the methodology of the original paper, we evaluate several AI safety guardrails, including Perspective API, HarmFormer, TTP with the original GPT-4o, our best-performing TTP model based on Llama 3.3 70B (Section 5.1), and LlamaGuard 3 (Inan et al., 2023), on the OpenAI Moderation Dataset (Markov et al., 2023). To ensure reproducibility and deterministic outputs, all generations are performed with a temperature of 0.

As for Perspective API, since Mendu et al. (2025) doesn't provide the specific threshold used, we test both the standard one (0.5) and the one optimized for TTP-eval (0.4).

Regarding the LlamaGuard models, we tried to reconstruct Mendu et al. (2025) original three prompting strategies: (1) *focused*, (2) *zero-shot*, and (3) *few-shot*. The *focused* prompt relies solely on the harm categories LlamaGuard 3 was originally trained on, serving as baseline. The *zero-shot* prompt adapts the system instructions to include OpenAI Moderation category definitions, while the *few-shot* prompt extends this by incorporating two detailed descriptions per category to guide the model in identifying harmful content according to the OpenAI taxonomy. Given the exact prompts were not provided in the original paper, we reconstructed these prompting strategies following the official Meta documentation. We based our implementation on the LlamaGuard 3 model card, and we directly utilized Meta's prompt formatting utilities in the Llama Cookbook repository.

Since LlamaGuard 3 is available in two parameter configurations (1B and 8B) and the original paper did not specify which version they evaluated, we test both variants across all three prompting strategies. Finally, given that each model was trained with a different harm taxonomy, we follow the Mendu et al. (2025) approach of aggregating all category-specific predictions into a single binary harmful/safe classification to enable fair comparison across models.

4.3 Web-scale Harm Analysis

Mendu et al. (2025) analyzed 3 million web pages sampled from Common Crawl (1M), C4 (1M), and FineWeb (1M) using their TTP prompt to quantify the prevalence of toxic and topical content in widely-used pre-training datasets. Given computational constraints and the prohibitive cost of running 3 million inferences, we conduct a scaled-down reproduction of their analysis using the three subsets of 10,000 pages described in Section 3.2. We apply our best-performing TTP model (Llama 3.3 70B, selected from Section 5.1) to classify each sampled webpage according to the three dimensional taxonomy across all harm categories. For each web page, the model produces:

- A dimension label (Safe, Topical, or Toxic) for each harm category.
- An aggregated toxic indicator: a webpage is considered Toxic overall if labeled Toxic in any of the harm categories.

- An aggregated topical indicator: a webpage is considered Topical overall if labeled Topical in any of the harm categories (excluding pages already marked as toxic).

Following the original methodology, we compute per-harm and aggregated statistics by calculating the percentage of web pages labeled as Toxic or Topical for each category and dataset.

4.3.1 Harmformer for web data analysis

To improve the significance of the TTP-based analysis on 10,000 pages per dataset, we extend our study by leveraging HarmFormer Mendu et al. (2025), a LongFormer (Beltagy et al., 2020) model fine-tuned by the original authors on a subset of their 3 million-page web corpus. Unlike TTP, which requires a full LLM inference per page, HarmFormer's encoder-only architecture enables substantially more efficient inference, making analysis at the scale of 1 million pages per dataset computationally feasible within our resource budget (see Section 6.2). Since HarmFormer is limited to a maximum context window of 1024 tokens, each webpage is first split into passages using Recursive Character Text Splitting, targeting a chunk size equal to the token budget and using a 200-character overlap between adjacent passages to preserve cross-boundary context. The resulting character-level chunks are then merged and tokenized before inference. Each passage is classified independently across the five harm categories and three taxonomy dimensions, and page-level labels are obtained via max-pooling: for each harm category, a page is assigned the highest harm score observed across any of its passages.

4.4 HAVOC: Harmful Abstractions and Violations in Open Completions Benchmark

HAVOC is a benchmark designed by Mendu et al. (2025) in order to assess a model's ability to avoid generating toxic completions when prompted with a toxic prefix. To construct the benchmark, Mendu et al. (2025) sampled 10,376 snippets from web pages, drawing from the 3 million candidate set previously established in Section 4.3. They employed two separate prompts to extract prefixes and suffixes from these snippets, although we noted that the suffix is ultimately unused in the evaluation pipeline, as confirmed by the original authors via correspondence. Since constructing the dataset from scratch was computationally infeasible, as it would require running TTP on the entire 3 Million samples dataset, we directly utilized the HAVOC benchmark dataset curated by the original paper. However, we observed that of the original 10,376 snippets reported by the original paper, 289 are effectively unusable, leaving the HAVOC benchmark dataset with only 10,087 snippets.

The primary scope of HAVOC is to measure the leakage rate of the model across harm categories. This is determined by analyzing system-generated completion to which the TTP attributes a toxic label against the prefixes labels. To assess this, TTP is run on both the input prefixes and the system-generated completions across several models.

Inference Configuration We generate 200-token completions for each HAVOC prefix using Gemma 2 (2B, 9B, 27B), Llama 3.2 (1B, 3B), and Mistral v0.3 7B. To match the original setup, we use the base (non–instruction-tuned) text models and run inference via Ollama with greedy decoding. Although the original methodology reports greedy decoding being used, the authors later confirmed that their experiments were conducted with a temperature of 0.8.

The second phase of the experiment required TTP to evaluate both the prefixes and the model generations in two separate runs.

Evaluation Protocol As per the original paper, each generation is classified into one of three categories based on the alignment between the prefix and completion labels:

- **Provocative Leak:** The prefix is *Toxic* and the generation is also labeled *Toxic*.
- **Passive Leak:** Occurs when the prefix is labeled as *Topical* but the completion is labeled as *Toxic*.
- **Neutral Leak:** Occurs when a *Toxic*-labeled prefix results in a completion labeled as *Safe*.

Intuitively, in the context of safety, a Neutral Leak is preferred, as it indicates the model is capable of generating safe responses even when prompted with toxic snippets.

5 Results

In this section we present the findings of our reproducibility study and compare them to the one presented by Mendu et al. (2025). We first analyze the adaptability of the TTP prompt on different open-source models evaluated over the TTP-eval dataset. Then after having selected the best-performing model we compare it along with other models on the OpenAI moderation dataset. We continue by using the best TTP model and HarmFormer to analyze the distribution of harmful content in the pretraining datasets and we end by reporting the HAVOC benchmarking findings.

5.1 TTP Prompt Evaluation on TTP-Eval

We successfully replicated the Perspective API baseline, obtaining a F1-score of 0.62. This result closely matches Mendu et al. (2025) reported performance. As can be observed in Table 1, our TTP evaluation experiments reveal both consistencies and notable discrepancies with the results reported in the original paper. For TTP, the original paper reports the highest F1-score of 0.83 using GPT-4o. In contrast, our greedy-decoding evaluation of GPT-4o yields an F1-score of 0.78, corresponding to a 5-point drop in F1 and a 10-point drop in recall. Increasing the temperature to 0.8 improves recall to 0.72 and raises F1 to 0.80, partially closing the gap. These discrepancies may arise from differences in model versions, decoding settings, or subtle prompt implementation details.

Mendu et al. (2025) claims that TTP performs poorly on non-GPT models such as DeepSeek-R1, Gemma 2, and Gemini 2, attributing this to prompt-tuning carried out specifically for GPT-4o. They report extremely low F1-scores for these models (e.g., DeepSeek-R1 32B: 0.06; Gemma 2 27B: 0.35). Our results directly contradict this claim. In particular:

- **DeepSeek-R1 32B** achieves F1 = 0.55, far above the reported 0.06.
- **Gemma 2 27B** reaches F1 = 0.67, nearly doubling the performance reported in the paper.
- **Gemini 2 Flash** reaches F1 = 0.75, 11 percentage points above the 0.64 reported in the original work.

Moreover, when TTP is evaluated on other open-weight models, the following results are noteworthy:

- **Llama 3.3 70B** achieves F1 = 0.80, matching GPT-4o despite having substantially fewer parameters.
- **Gemma 3 27B** achieves F1 = 0.77 with balanced precision and recall.

Our main findings are the following:

- Strong performances are not limited to proprietary models. Llama 3.3 70B achieves the highest F1-score across all models that we tested, while offering full reproducibility and no API costs.
- Model size is not the sole determinant of performance. Gemma 3 27B achieves an F1-score of 0.77 while fitting on a single 40GB GPU, providing an attractive efficiency–performance trade-off and enabling reproducible evaluation in resource-constrained settings.
- Models optimized for chain-of-thought reasoning do not necessarily perform well on this task. Despite their design for complex reasoning, DeepSeek-R1 models underperform relative to similarly sized standard models.
- We observed that models frequently struggled to adhere to the TTP prompt's required output format. Smaller models ($<$ 30B parameters) exhibited more formatting errors, but even larger models occasionally produced malformed outputs.

Family	Model	P	R	F1
GPT	GPT-4o	0.89	0.69	0.78
	GPT-4o (t=0.8)	0.89	0.72	**0.80**
	GPT-OSS 20B	0.86	0.54	0.67
	GPT-OSS 120B	0.71	**0.77**	0.74
DeepSeek	DeepSeek-R1 70B	0.80	0.54	0.65
	DeepSeek-R1 32B	0.69	0.46	0.55
Qwen	Qwen2.5 72B	0.88	0.67	0.76
	Qwen3 32B	0.74	0.70	0.72
Llama	Llama 3.3 70B	0.85	0.76	**0.80**
	Llama 3.3 (t=0.8)	0.86	0.67	0.75
	Llama 3.3 (Q3)	**0.90**	0.65	0.75
Mistral	Mistral L. 123B	0.85	0.67	0.75
	Mixtral 8x7B	0.85	0.64	0.73
Gemma/Gemini	Gemma 2 27B	0.79	0.57	0.67
	Gemma 3 27B	0.77	**0.77**	0.77
	Gemini 2 Flash	0.87	0.66	0.75

Table 1: TTP prompt performance on TTP-Eval. A webpage is Toxic if any harm category is detected. Bold indicates best and underline second-best performance per metric.

To further test the hypothesis of GPT-specific prompt-tuning, we evaluate GPT-OSS models (20B and 120B). Despite being developed by OpenAI and sharing some training techniques with GPT-4o, these models achieve F1-scores of only 0.67–0.74 and do not match the performance of Llama 3.3 70B.

5.2 Comparison with OpenAI Moderation Dataset

Table 2 presents the performances of the evaluated safety guardrails on the openAI moderation dataset. Our analysis highlights several key findings regarding the effectiveness of open-source models compared to those of Mendu et al. (2025). We observe a near exact alignment between our reproduction and the original reported values for established baselines. Specifically, our evaluation of HarmFormer (F1: 0.73) and Perspective API at threshold 0.5 (F1: 0.66) matches the original paper's scores identically. This confirms that our evaluation pipeline and metric calculation are consistent with the original methodology. A critical finding of our study addresses an ambiguity in the original text regarding the LlamaGuard (LG) version employed. Mendu et al. (2025) reported F1 scores (F1: 0.66-0.70) for LlamaGuard 3. Our experiments show that these figures align closely with the 1B parameter variant (F1: 0.66-0.68). In contrast, we found that the 8B parameter variant significantly outperforms the original reported baselines, achieving an F1 score of 0.80.

This suggests the original study likely defaulted to the smaller model, thereby underestimating the potential of the bigger version compared to their proposed TTP method. Moreover, our results demonstrate that the effectiveness of the TTP prompt is not exclusive to proprietary models. While the original paper reported an F1 score of 0.80 using GPT-4o, our open-source implementation using Llama 3.3 70B achieved a comparable F1 score of 0.81. Furthermore, our reproduction of the TTP (GPT-4o) pipeline yields a slightly higher F1 score (0.83) than originally reported, possibly due to minor variances in API behavior or prompt handling. Overall, this validates that Llama 3.3 70B is a sufficient and robust alternative for implementing the Mendu et al. (2025) safety taxonomy without relying on closed source APIs.

5.3 Prevalence of Toxic & Topical Text in Common Crawl Sets found by TTP

Our analysis on 30,000 webpages (10,000 per dataset) using TTP with Llama 3.3 70B reveals both consistencies and divergences from the original findings, as shown in Table 3.

Toxic Content We observe strong agreement on toxic content prevalence across all datasets. Common Crawl (4.2% vs. 4.1% original), C4 (1.7% vs. 2.1%), and FineWeb (2.9% vs. 3.9%) differ by at most 1.0 percentage point. Sexual harm dominates in Common Crawl (2.2%), confirming the original finding, while ideological harm is most prevalent in C4 and FineWeb, reflecting their curation strategies.

	Reproduction			Original		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Perspective (0.5)	0.82	0.55	0.66	0.82	0.55	0.66
LG 3.3 1B Focused	0.55	0.81	0.66	0.56	0.81	0.66
LG 3.3 1B Zero	0.58	0.83	0.68	0.62	0.75	0.68
LG 3.3 1B Few	0.60	0.78	0.68	0.64	0.78	0.70
TTP (GPT-4o)	0.75	**0.92**	**0.83**	0.72	0.91	0.80
HarmFormer	0.64	0.84	0.73	0.64	0.84	0.73
Perspective (0.4)	0.76	0.66	0.71	–	–	–
LG 3.3 8B Focused	0.81	0.76	0.79	–	–	–
LG 3.3 8B Zero	**0.86**	0.75	0.80	–	–	–
LG 3.3 8B Few	**0.86**	0.73	0.79	–	–	–
TTP (Llama 70B)	0.74	0.90	0.81	–	–	–

Table 2: Performance on the OpenAI moderation dataset. Rows with dashes correspond to models or settings not reported in Mendu et al. (2025). Bold indicates best and underline second-best performance per metric.

Topical Content We consistently identify higher topical content rates than originally reported: Common Crawl (6.2% vs. 3.4%), C4 (13.4% vs. 6.8%), and FineWeb (18.1% vs. 10.6%). This systematic divergence is primarily driven by Ideological Harm classifications, likely reflecting model-specific differences in interpreting the Safe-Topical boundary between Llama 3.3 and GPT-4o, or sampling variation between crawl snapshots.

Dataset	Dimension	Hate & Violence	Ideological	Sexual	Illegal	Self-Inflicted	Total
Common Crawl	Toxic	0.3%	0.8%	**2.2%**	0.8%	0.6%	4.2%
	Topical	1.1%	**3.4%**	1.0%	1.5%	0.7%	6.2%
C4	Toxic	0.4%	**0.8%**	0.3%	0.4%	0.2%	1.7%
	Topical	3.7%	**8.0%**	1.2%	2.2%	1.5%	13.4%
FineWeb	Toxic	0.6%	**1.4%**	1.0%	0.4%	0.1%	2.9%
	Topical	5.5%	**11.1%**	2.7%	2.7%	1.8%	18.1%

Table 3: Labeled harm prevalence (%) across pretraining datasets by category.

5.3.1 Key Findings and Interpretations

Toxic Content Prevalence Our results strongly support the original paper's central claim that pretraining datasets contain non-negligible amounts of toxic content, with measurements within 0.5 percentage points for most categories. Sexual harm dominates toxicity in Common Crawl (53% of total toxic content, compared to 78% reported in the original study), while ideological harm is the largest category in C4 and FineWeb, with sexual harm reduced to 15% for C4 and 34% for FineWeb, confirming the differential effects of dataset curation strategies.

Topical Content Divergence We consistently found 1.7-2.0× higher topical content rates than originally reported. This systematic divergence likely stems from model-specific biases in interpreting the Safe-Topical boundary. Our Llama 3.3 70B may classify borderline cases as Topical more readily than GPT-4o, or conversely, GPT-4o may be more conservative and biased toward Safe classifications. Particularly for Ideological Harm (3.4% vs. 0.7% in CC, 8.0% vs. 1.6% in C4), heightened sensitivity to content discussing conspiracy theories and misinformation might lead to systematic higher values in our findings, a pattern also observed in our HAVOC results (Section 5.5). The strong agreement on toxic content ($< 0.5\%$ differences) confirms the core safety claims are reproducible, while the topical divergence suggests this dimension warrants additional scrutiny in content curation applications.

Dataset-Specific Patterns Our analysis confirms the original Mendu et al. (2025) observation that dataset curation strategies have differential effects on toxic versus topical content:

- **C4's Keyword Filtering**: Effectively reduces Sexual toxicity (0.3% vs. 2.2% in Common Crawl) by removing explicit pornographic content. However, it preserves educational sexual health discussions, resulting in comparable or slightly higher topical rates (1.2% vs. 1.0%).
- **Quality vs. Safety Trade-offs**: Both C4 and FineWeb show higher topical content rates than toxic content rates, suggesting that quality-focused filtering preserves substantive discussions of sensitive topics while removing overtly harmful material.
- **Ideological Content Prevalence**: The high topical rates for Ideological Harm across all datasets highlight the challenge of distinguishing legitimate discussion of controversial topics from harmful misinformation or bias, a key motivation for the three dimensional taxonomy.

Statistical Validation To assess the statistical significance of observed differences, we conducted several Two-Sample Z-tests for proportions, comparing our findings across three datasets (CC, C4, and FW) against the baseline established by Mendu et al. (2025). As detailed in Table 4, the majority of comparisons yielded significant deviations ($p < 0.05$), particularly within the Topical dimension, where our setting (model and sub-sampling) consistently captured a higher percentage of relevant content. Notably, the Toxic dimension for the CC dataset was the only instance where we failed to reject the null hypothesis ($Z = 0.5, p = 0.61$), indicating that our results for this specific subset are in line with the one of the original report.

Dataset	Dimension	Paper %	Our %	Z-score	p-value	Non-Sig. Range	Significant
CC	Toxic	4.1%	4.2%	0.5	0.61	[3.71 - 4.48]	×
	Topical	3.4%	6.2%	15.3	0.0	[3.04 - 3.75]	✓
C4	Toxic	2.1%	1.7%	-2.78	0.005	[1.82 - 2.38]	✓
	Topical	6.8%	13.4%	25.97	0.0	[6.31 - 7.29]	✓
FW	Toxic	3.9%	2.9%	-5.15	0.0	[3.52 - 4.28]	✓
	Topical	10.6%	18.1%	24.17	0.0	[10.00 - 11.28]	✓

Table 4: Statistical comparison of our results with the original paper. "Significant" indicates the null hypothesis is rejected at $\alpha = 0.05$.

Implications for Reproducibility The strong alignment in toxic content measurements (differences $<$ 0.5%) demonstrates that the core claims of the original paper regarding harmful content prevalence are reproducible using open-source models and publicly available datasets. However, the substantial topical content divergence (differences of $\sim 2.0\times$) highlights a key limitation: the boundary between Safe and Topical classifications appears sensitive to model choice and implementation details.

5.4 Toxic & Topical Text analysis in Common Crawl Sets by HarmFormer

Our extended analysis leverages the HarmFormer model proposed by Mendu et al. (2025) to classify harmful content at a scale of 1 million webpages per dataset, substantially exceeding the 10,000 pages used in our TTP analysis (Section 4.3). Given HarmFormer's efficient inference on fixed-length passages, this larger sample allows for a more statistically robust characterization of harm prevalence, as summarized in Table 5.

Toxic Content HarmFormer consistently detects higher toxic content rates than both the original TTP findings and our own TTP reproduction. Common Crawl reaches 11.1% (vs. 4.1% in the original and 4.2% in our TTP run), C4 reaches 3.9% (vs. 2.1% original), and FineWeb 5.6% (vs. 3.9% original). Notably, in Common Crawl, Sexual Content and Illegal Activities each account for 4.6% of toxic content, jointly driving the elevated total, whereas our TTP analysis identified Sexual harm alone (2.2%) as the dominant category. This discrepancy likely reflects architectural differences: HarmFormer, a fine-tuned LongFormer operating on passage-level representations, may be more sensitive to distributional signals indicative of harm than a prompt-based classifier.

Topical Content Topical content rates found by HarmFormer are broadly consistent with our TTP results, particularly for Common Crawl (6.1% vs. 6.2% with TTP) and FineWeb (16.1% vs. 18.1%). C4 shows a moderate difference (10.5% vs. 13.4%), but both substantially exceed the original paper's reported 6.8%. This convergence between TTP and HarmFormer on the Topical dimension lends additional confidence to our earlier finding that topical content is systematically underreported in Mendu et al. (2025).

Statistical Validation As shown in Table 6, all comparisons against the original paper yield statistically significant deviations ($p < 0.001$, Z-scores ranging from 56 to 186), without exception. This stands in contrast to our TTP analysis, where CC Toxic was the sole non-significant result. The significance is partly a consequence of the larger sample ($n = 1$M), which substantially narrows confidence intervals, but the magnitude of the effect sizes, particularly for CC Toxic ($Z = 186.8$), indicates genuine distributional differences attributable to the classifier rather than sampling alone.

Implications for Classifier Choice The systematic divergence between HarmFormer and TTP on toxic content (e.g., 11.1% vs. 4.2% for CC) highlights a key sensitivity of web-scale harm estimates to classifier architecture. HarmFormer's passage-level fine-tuning may surface harm signals that prompt-based models de-emphasize, or it may suffer from false positive accumulation across the multi-passage max-pooling aggregation strategy. While both classifiers agree qualitatively, confirming that pretraining datasets contain non-negligible harmful content, quantitative estimates diverge substantially on the Toxic dimension, underscoring the need for multi-classifier validation in safety-critical data curation pipelines.

Dataset	Dimension	Hate & Violence	Ideological	Sexual	Illegal	Self-Inflicted	Total
Common Crawl	Toxic	0.5%	0.9%	4.6%	4.6%	1.3%	11.1%
	Topical	2.3%	1.7%	1.2%	1.3%	0.9%	6.1%
C4	Toxic	0.4%	1.0%	0.8%	1.4%	0.6%	3.9%
	Topical	4.9%	3.1%	1.1%	1.7%	1.3%	10.5%
FineWeb	Toxic	0.9%	1.4%	1.9%	1.3%	0.8%	5.6%
	Topical	7.5%	4.9%	2.8%	2.2%	1.8%	16.1%

Table 5: Labeled harm prevalence (%) across pretraining datasets by category, as classified by HarmFormer on 1 million samples per dataset.

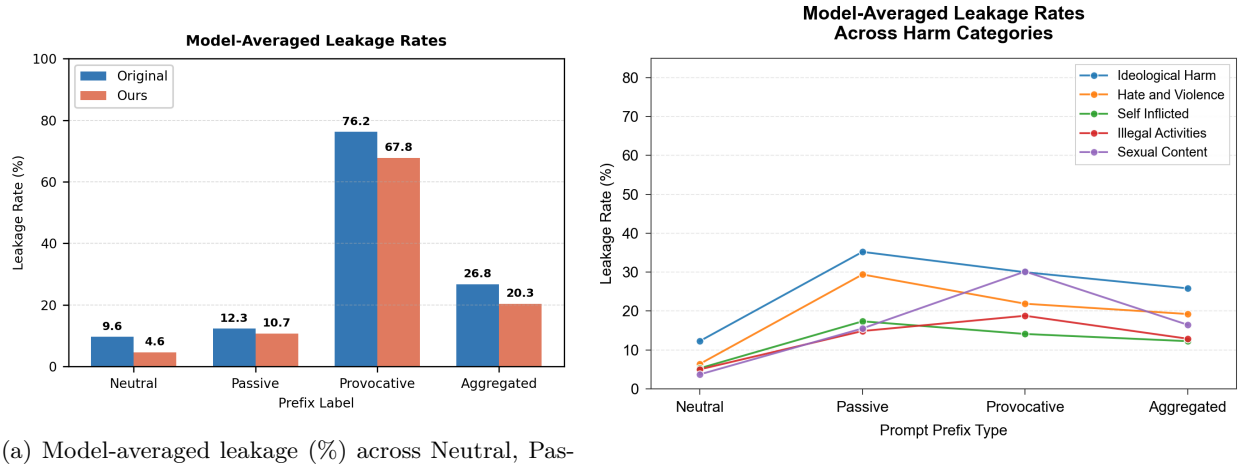
Dataset	Dimension	Paper %	Our %	Z-score	p-value	Non-Sig. Range	Significant
CC	Toxic	4.1%	11.1%	186.79	0.0	[4.06, 4.14]	✓
	Topical	3.4%	6.1%	89.76	0.0	[3.36, 3.44]	✓
C4	Toxic	2.1%	3.9%	74.61	0.0	[2.07, 2.13]	✓
	Topical	6.8%	10.5%	93.07	0.0	[6.75, 6.85]	✓
FW	Toxic	3.9%	5.6%	56.51	0.0	[3.86, 3.94]	✓
	Topical	10.6%	16.1%	114.35	0.0	[10.54, 10.66]	✓

Table 6: Statistical comparison of HarmFormer results (1M samples per dataset) against the original paper Mendu et al. (2025). All comparisons are significant at $\alpha = 0.05$. "Significant" indicates the null hypothesis is rejected.

5.5 HAVOC: Harmful Abstractions and Violations in Open Completions Benchmark

Our HAVOC benchmark evaluation reveals that open-source text completion models exhibit consistent toxicity leakage patterns across model families and parameter scales. While we successfully reproduce the general trend reported by Mendu et al. (2025), as shown in Figure 1a, our results indicate a Provocative leak rate averaging 67.8% across all tested models, which is notably lower than the 76% reported in the original study. Furthermore, the leakage rate remains relatively stable across different model architectures and sizes, suggesting that toxicity leakage is driven primarily by pre-training data rather than model capacity. This finding underscores the importance of dataset filtering as a foundational intervention for improving model safety.

Another key divergence from the original paper concerns the distribution of harm types, as illustrated in Figure 1b. While Mendu et al. (2025) identified Sexual Harm as the most prevalent category in their analysis, our evaluation instead finds Ideological Harm to be the dominant source of toxic completions across the HAVOC benchmark. The web-scale analysis presented in Section 4.3 further corroborates this pattern, with Ideological Harm exhibiting the largest deviation from the original findings of Mendu et al. (2025) and consistently appearing at higher rates across all our experiments. This discrepancy may reflect differences in the specific model versions evaluated, as well as variation in how different model families handle ideologically sensitive content as part of their pre-training and training process.



(a) Model-averaged leakage (%) across Neutral, Passive and Provocative tones on the HAVOC dataset.

(b) Model Averaged leakage plot of HAVOC across harms.

Figure 1: Comparison of model-averaged leakage on HAVOC with Mendu et al. (2025).

6 Discussion

6.1 Interpretation of Key Divergences

Our comparative analysis reveals a systematic and statistically significant divergence in the classification of Topical content: Llama 3.3 consistently identifies $\sim 2.0\times$ more topical material than the GPT-4o baseline reported by Mendu et al. (2025), with the most pronounced differences concentrated in the Ideological Harm category (e.g., 3.4% vs. 0.7% in Common Crawl). Crucially, the near-identical topical rates produced by our two independent classifiers, TTP with Llama 3.3 (6.2%) and HarmFormer (6.1%) for Common Crawl, suggest that this divergence is not a Llama specific artifact. Rather, it points to a fundamental sensitivity of the Safe-Topical boundary to model choice, with the convergence of two independent classifiers suggesting the divergence from the original is systematic rather than incidental. In general, our experiments demonstrate that prompt based safety taxonomies cannot be treated as model-agnostic solutions: practitioners should validate classifier agreement on a representative subset before committing to full scale deployment.

Comparing our two web-scale evaluation approaches exposes a clear contrast in toxic content estimation. HarmFormer detected substantially higher toxicity rates in Common Crawl (11.1%) compared to our TTP-based analysis (4.2%). One plausible structural explanation is that HarmFormer's passage level max-pooling aggregation (where a document is flagged as toxic if any passage exceeds the harm threshold) accumulates false positives over long documents in a way that full document TTP evaluation does not. However, we did not conduct a passage level error analysis to verify this hypothesis directly, and architectural differences in sensitivity between a fine-tuned LongFormer and a prompt-based LLM may independently contribute to the gap. Both classifiers agree qualitatively on the rank ordering of datasets (CC > FineWeb > C4) and on the Topical dimension, which reinforces the core reproducibility claim while underscoring that absolute toxic content rates are classifier-dependent.

6.2 Computational Requirements

All experiments were conducted on the Snellius supercomputer operated by SURFsara, constrained to at most two NVIDIA A100 40GB GPUs to ensure sustainable resource usage across large-scale experiments. This budget directly drove model selection: Llama 3.3 70B, distributed across two A100s, provided the best performance/cost ratio (F1 = 0.80) among the evaluated open-weight models. In terms of API costs, a single TTP evaluation run on GPT-4o amounted to approximately $23.00 on the OpenAI Moderation Dataset alone, compared to near-zero marginal cost for Llama 3.3 70B, further motivating our open-weight approach.

The most striking computational contrast in our study is between the two web-scale classifiers. TTP with Llama 3.3 70B processed 30,000 webpages in approximately 106 GPU hours (∼3.5 seconds per page), while HarmFormer classified 3 million webpages using a single A100 in only 33 GPU hours (∼0.04 seconds per page), this led to a 100× increase in throughput at one-third the GPU cost. This difference has direct implications for practitioners: TTP offers higher interpretability and flexibility, but HarmFormer is the only practically viable option for safety auditing at true web-scale. Total carbon emissions across all experiments were estimated at 45.50 kg CO_2 equivalent (≈ 184 km driven by an average ICE vehicle). Full per-experiment breakdowns of GPU hours, VRAM usage, wall-clock time, and API costs are reported in Appendix A.1.

6.3 Challenges

The reproducibility effort faced significant hurdles, primarily because the code of the original work was not provided, requiring a complete reimplementation of the framework from scratch. This process was further complicated by methodological gaps and inconsistencies in the original works, such as undisclosed sampling procedures, missing prompt details, and ambiguities regarding model sizes and dataset counts. Furthermore, the prohibitive costs associated with the original proprietary models (like GPT-4o) necessitated a scaled-down analysis and a shift toward computationally feasible open-source alternatives.

7 Conclusions

This work reproduces and extends the analysis of the three-dimensional safety taxonomy proposed by Mendu et al. (2025). We show that the taxonomy is reproducible with open-weight models alone and that this reproduction uncovers previously unreported limitations. Llama 3.3 70B matches GPT-4o on TTP-Eval at a fraction of the cost, while our web-scale analysis shows that the Safe/Topical boundary is systematically model-dependent, with Ideological Harm rates diverging by up to 5× across classifiers. The 2.6× discrepancy in toxic content estimates between TTP and HarmFormer further demonstrates that absolute harm rates in pretraining datasets are not a stable quantity, but rather are a function of classifier architecture. Practitioners deploying safety pipelines at scale should treat single-classifier estimates with caution and validate cross-classifier agreement before committing to filtering thresholds.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Common Crawl Foundation. Common crawl - open repository of web crawl data, 2024. URL https://commoncrawl.org/. Accessed: 2026-04-20.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. Safeguarding large language models: a survey. *Artificial Intelligence Review*, 58, 2025.

Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. Integrating content moderation systems with large language models. *ACM Trans. Web*, 19(2), 2025. ISSN 1559-1131. doi: 10.1145/3700789. URL https://doi.org/10.1145/3700789.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts, 2024. URL https://arxiv.org/abs/2404.05993.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning, 2019. URL https://arxiv.org/abs/1910.09700.

LangChain. URL https://github.com/langchain-ai/langchain. Accessed: 2026-04-20.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539147. URL https://doi.org/10.1145/3534678.3539147.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Mariani, Nicola von Werra, Gautier Malartic, and et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021. doi: 10.18653/v1/2021.emnlp-demo.21.

Llama Cookbook repository. URL https://github.com/meta-llama/llama-cookbook. Accessed: 2026-04-20.

LlamaGuard 3 model card. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-3/. Accessed: 2026-04-20.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world, 2023. URL https://arxiv.org/abs/2208.03274.

Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. Towards safer pre-training: analyzing and filtering harmful content in webscale datasets for responsible llms. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/53. URL https://doi.org/10.24963/ijcai.2025/53.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models, 2025. URL https://arxiv.org/abs/2506.04689.

Nowtricity. Carbon intensity of electricity, netherlands. URL https://www.nowtricity.com/country/netherlands. Accessed: 2026-04-20.

Ollama. URL https://ollama.com/. Accessed: 2026-04-20.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL https://doi.org/10.1145/3351095.3372873.

U.S. Environmental Protection Agency. Greenhouse gas equivalencies calculator - calculations and references. URL https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator-calculations-and-references. Accessed: 2026-04-20.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 1–18, Barcelona, Spain, January 2020. ACM. ISBN 9781450369367. doi: 10.1145/3351095.3372833.

A Appendix

A.1 Computational Requirements

The implementation of the TTP prompt at scale presented substantial computational challenges, as TTP represents a single inference operation evaluating one webpage across the three-dimensional taxonomy, while HAVOC applies this mechanism at scale by requiring TTP evaluations for both input prefixes and generated completions across thousands of samples and multiple models. Consequently, model selection for the optimal TTP classifier (Section 5.1) was constrained by computational feasibility, with a target of at most two NVIDIA A100 40GB GPUs to ensure sustainable resource usage for large-scale experimentation, ultimately leading to the selection of Llama 3.3 70B deployed via Ollama on this hardware configuration for all subsequent experiments.

TTP-Eval Benchmark Processing the TTP-Eval dataset (393 webpages) with Llama 3.3 70B required 1 hour and 8 minutes of wall-clock time and consumed 2,319,600 tokens (2,246,790 input; 72,810 output). Across all open-weight models, the median processing time was 11 seconds per webpage. Evaluating 13 open-source model variants in TTP-Eval required approximately 47.6 GPU hours, while for proprietary baselines, a single TTP-Eval run cost approximately $5.00 for GPT-4o and $0.257 for Gemini 2 Flash.

OpenAI Moderation Dataset Benchmark Evaluating AI safety guardrails on the OpenAI Moderation Dataset (1,680 samples) presented variable computational requirements depending on model architecture. Perspective API, operating via CPU with a rate limit of approximately 1 request/s, required 30 minutes of inference time, while TTP using GPT-4o via API incurred costs of approximately $23.00[3] per evaluation run. HarmFormer executed on GPU MIG in approximately 3 minutes, LlamaGuard variants required 10 minutes (1B model) and 15 minutes (8B model) per prompting configuration on an A100 GPU, and TTP with Llama 3.3 70B required approximately ~ 5 GPU hours for complete dataset evaluation.

Web-Scale Harm Analysis Processing 30,000 webpages with TTP on Llama 3.3 70B required substantial computational resources: Common Crawl consumed 39 GPU hours, C4 32 GPU hours, and FineWeb 35 GPU hours, totalling 105.88 GPU hours across two A100s (see Table 8). By contrast, the HarmFormer-based analysis of 3 million webpages required only a single A100 GPU, yielding 33.25 total GPU hours despite processing 100× more samples. The stark contrast between the two pipelines, roughly 3.5 seconds per page for TTP versus 0.04 seconds per page for HarmFormer, underscores the practical scalability benefits of fine-tuned lightweight classifiers for safety-critical data curation at web-scale.

HAVOC Benchmark The HAVOC experiment required separate evaluations for both prefix classification and completion generation assessment across 10,087 samples per model. Text generation varied by model size: smaller models (1B-3B parameters) required ∼15-20GB VRAM and 4-5 hours, while larger models (9B-27B parameters) required full 40GB allocation and 8-10 hours, totaling ∼52 GPU hours. The subsequent TTP evaluation phase, using Llama 3.3 70B to classify both prefix and completion columns, required approximately 32 GPU hours per column, totaling 224 GPU hours for the classification component alone.

A.2 Environmental impact

All experiments were conducted on the Snellius supercomputer operated by SURFsara in the Netherlands. Emissions were estimated using the MachineLearning Impact calculator introduced by Lacoste et al. (2019). Assuming an average carbon intensity of the Dutch electricity grid of 0.388 $kgCO_2$ eq/kWh (Nowtricity), the total consumption of our experiments amounted to 45.50 $kgCO_2$ equivalent, corresponding to approximately 184 km driven by an average internal combustion engine (ICE) car (U.S. Environmental Protection Agency).

[3]This represents an upper bound as it does not account for prompt caching.

A.3 Additional Figures

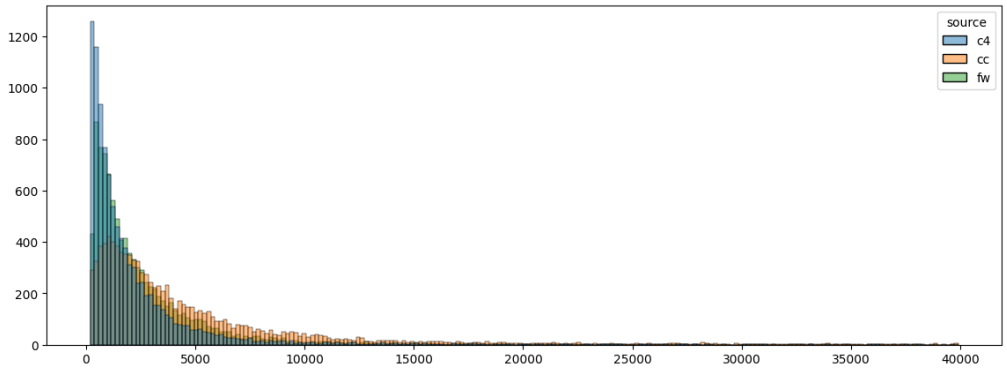


Figure 2: Distribution of document lengths for Common Crawl (CC), C4, and FineWeb. The figure shows the distribution of document lengths after truncation, retaining only documents between 200 and 40,000 characters. Despite truncation, the distributions remain strongly long-tailed, indicating that a small number of very long documents continue to inflate the mean document length.

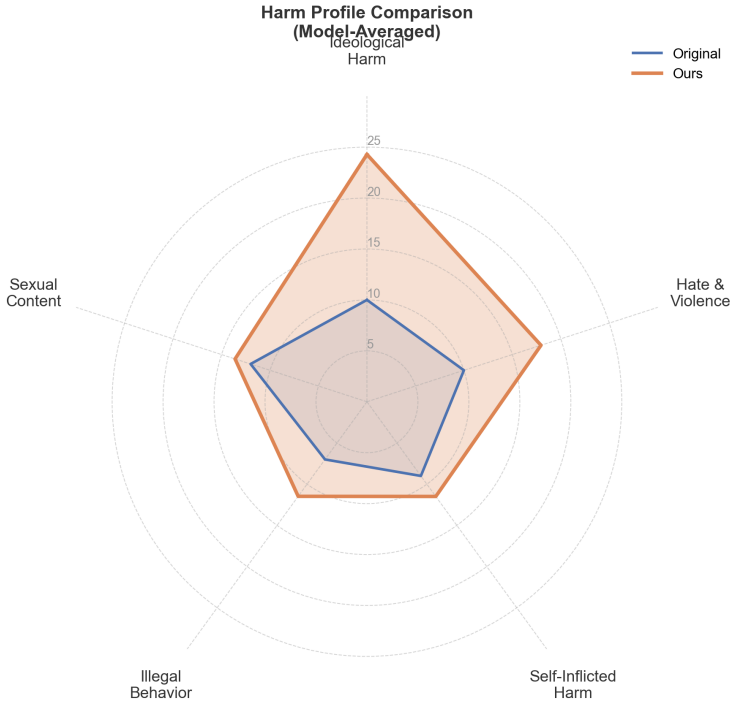


Figure 3: Comparison of harm detection results with original HAVOC findings. The figure presents a comparison between our replication study and the original authors' findings regarding non-safe content, including both topical and toxic harm percentages. Note that samples may belong to multiple harm categories simultaneously.

A.4 Additional Tables

Dataset	Range		Before Filtering			After Filtering		
	Min	Max	Mean	Std	Median	Mean	Std	Median
CC	0	452,805	5489.41	3362.99	1151.0	4844.39	5698.91	2955.0
C4	36	69,815	2133.42	13205.24	2774.0	2176.02	2998.18	1222.5
FW	136	119,484	3050.40	5202.58	1759.0	2809.18	3509.41	1750.0

Table 7: Statistics of document length distributions for web-scale datasets. The table compares document length statistics (in characters) for Common Crawl (CC), C4, and FineWeb (FW) before and after applying length-based filtering.

Dataset	Samples	Wall Time	Sec/Sample	GPU Hours
TTP with Llama 3.3 70B (2× A100)				
Common Crawl	10,000	19:30:30	7.02	39.02
C4	10,000	15:54:10	5.72	31.81
FineWeb	10,000	17:31:23	6.31	35.05
Total	30,000	2 days, 4:56:03	6.35	105.88
HarmFormer (1× A100)				
Common Crawl	1,000,000	17:29:08	0.063	17.49
FineWeb	1,000,000	8:23:19	0.030	8.39
C4	1,000,000	7:22:16	0.027	7.37
Total	3,000,000	1 day, 9:14:43	0.040	33.25

Table 8: Computational costs for web-scale harm analysis. Wall time represents total elapsed time per dataset. GPU hours for TTP are calculated as wall-clock time multiplied by 2 GPUs; HarmFormer ran on a single A100.

Experiment	GPU Hours	CO_2 Emissions (kg)
TTP	47.6	4.62
Safety guardrails comparison	6.15	0.59
Web-scale analysis (10K)	105.88	10.27
Web-scale analysis (1M)	33.25	3.22
HAVOC	276.3	26.8

Table 9: Environmental impact of computational experiments.