‘Mamma Mia!’ 👌 Becomes ‘Mamma Mild’ 🤗: A Benchmark Dataset and Model Evaluation for Meaning-Preserving Detoxification of Italian Social-Media Comments

Anonymous ACL submission

Abstract

The challenge of detoxifying social media comments without altering their original meaning is particularly acute for non-English languages like Italian, where standard models often distort content. To address this, we introduce *MammaMild*, the first benchmark dataset for meaning-preserving detoxification in Italian, containing 5,000 toxic comments and their human-validated, detoxified counterparts designed to preserve the original meaning and structural properties. The dataset was constructed using a human-in-the-loop process with a top-performing LLM (Gemini) providing initial rewrites. We then fine-tuned six open-source models (1B–24B parameters) on *MammaMild* and developed WAS, a novel metric that jointly evaluates toxicity reduction, semantic fidelity, and length fidelity. Our experiments show that fine-tuning yields substantial improvements, a finding confirmed by both automated metrics and blind human evaluations. Finally, we demonstrate that the dataset can be repurposed to train highly accurate classifiers for distinguishing between toxic and optimally detoxified text. Our dataset, models, and code are publicly available.

ORIGINAL TOXIC

”Dio bono, ragioni come uno scemo deficiente. Mi hanno rotto i coglioni gli idioti come te che danno pareri senza avere la minima di idea di ciò di cui parlano.”

(English: ”Good God, you reason like a dumb idiot. I’m so pissed off by these morons like you that give opinions and have no clue what they’re talking about.”)

DETOX (Non-Meaning-Preserving)

”Manteniamo la discussione su toni rispettosi.”

(English: ”Let’s have a respectful discussion.”)

DETOX(Meaning-Preserving)

”Santo cielo, ragioni proprio come uno stupido di prima categoria. Non ne posso davvero più dei sempliciotti come te che danno pareri senza avere la minima di idea di ciò di cui parlano.”

(English: ”Gosh, you reason just like a first-class fool. I really can’t stand simpletons like you giving opinions without having the slightest idea of what they’re talking about.”)

Figure 1: Example of text detoxification, illustrating meaning-preserving and non-meaning-preserving approaches.

1 Introduction

The widespread presence of toxic content on social media poses significant risks to online discourse and user well-being (Müller and Schwarz, 2021). While simple deletion raises censorship concerns (Habibi et al., 2024), a promising alternative is text detoxification: rewriting content to reduce toxicity while preserving its original meaning and style (Laugier et al., 2021; Dale et al., 2021). Crucially, detoxification that fails to preserve meaning is not an alternative to censorship, but merely a different form of it, silencing a user’s intent rather than their words. Figure 1 exemplifies this crucial distinction between non-meaning-preserving and meaning-preserving detoxification. This task, however, remains a major challenge. Even state-of-the-art Large Language Models (LLMs), despite their linguistic prowess, often struggle with this nuance; their alignment training can cause them to refuse toxic inputs or produce overly sanitized outputs that distort the original message (Zhang et al., 2024; He et al., 2024). This problem is severely amplified for languages like Italian, where the scarcity of high-quality datasets has hindered the development and evaluation of reliable detoxification models (Luo et al., 2024; Logacheva et al., 2022).

This leads to our primary research question:

> **Problem Statement**
>
> How can we automatically detoxify Italian social media comments by optimally reducing toxicity to the greatest extent possible without altering their original meaning and intent?

We break this problem down into several specific research questions: **RQ1:** How well do current state-of-the-art LLMs perform meaning-preserving detoxification in Italian compared to human rewrites? **RQ2:** Can LLMs or ML classifiers based on automated metrics reliably emulate human judgments regarding optimal detoxification and semantic equivalence? **RQ3:** Does fine-tuning on a specialized Italian detoxification dataset significantly enhance detoxification quality and meaning retention, as assessed by both automated metrics and human evaluation? **RQ4:** How can meaning preservation and toxicity reduction be balanced in evaluation? Can a new metric effectively measure both aspects, and what insights do its results and human assessments provide about trade-offs? **RQ5:** Can the *MammaMild* dataset be repurposed to train effective binary classifiers that distinguish toxic comments from their optimally detoxified counterparts?

Our Approach. We present *MammaMild*, a 5,000-pair dataset of toxic Italian comments and their meaning-preserving rewrites. Our creation process involved a human-in-the-loop workflow: after benchmarking leading LLMs, we used the top performer (Gemini) for initial generation, followed by rigorous validation from human experts. This high-quality corpus enabled us to fine-tune powerful Italian detoxification models and train accurate classifiers, confirming their effectiveness through both automated and human evaluation.

Contribution. Our key contributions are: (1) *MammaMild*, the first high-quality Italian dataset for meaning-preserving detoxification, sourced from comments across a diverse range of social media platforms and accompanied by a detailed linguistic analysis; (2) a comprehensive analysis of state-of-the-art models for both detoxification performance and their ability to mimic human quality judgments, complemented by an in-depth qualitative analysis of their linguistic behavior and failure modes; (3) fine-tuning open-source language models, including an ablation study demonstrating the dataset's high data efficiency; (4) validating the dataset's utility on a new downstream task by training highly accurate binary classifiers that can distinguish between toxic and optimally detoxified content; (5) introducing a novel metric (WAS) that offers a holistic evaluation by balancing toxicity reduction with semantic and length fidelity; and (6) publicly releasing our dataset, models, and evaluation code.

Paper Organization. We review related work (§2), describe the dataset's creation (§3) and analysis (§4), present our experiments (§5), and conclude (§6).

2 Related Work

Detoxification aims to rewrite toxic text into a neutral variant while preserving its original meaning (Laugier et al., 2021; Dale et al., 2021; Hallinan et al., 2023). This is related to the ongoing challenge of detecting hate speech, where research has been extensive but often limited to English (Hartvigsen et al., 2022; Patel and Alsobeh, 2025; Saha et al., 2022; Yimam et al., 2024).
Early detoxification work by Dale et al. (2021) used controlled generation to improve semantic preservation over older style-transfer methods (Guo et al., 2023). More recently, MaRCo was introduced to handle subtle toxicity (Hallinan et al., 2023). While large language models (LLMs) have shown detoxification capabilities (Shaikh et al., 2022), alignment tuning can cause them to excessively sanitize or refuse toxic inputs (Zhang et al., 2024). To address this, systems like DetoxLLM incorporate rationales (Khondaker et al., 2024), and lightweight prompt-based methods like YOPO aim to maintain semantic fidelity, though sometimes with residual toxicity (He et al., 2024).
Research remains heavily focused on English, with large parallel corpora like ParaDetox (Logacheva et al., 2022) being rare for other languages. Consequently, multilingual detoxification often relies on less precise translation-based methods (Luo et al., 2024). In Italian, the linguistic use of slurs has been widely studied (Bianchi, 2017; Cepollaro, 2015; Cepollaro and Labinaz, 2019; Cepollaro et al., 2019, 2024; Cruschina and Gianollo, 2024; Goffredo et al., 2022; Klein, 2018), but detoxification research is scarce, with only a single bench-

mark available (De Ruvo et al., 2025). Our work, *MammaMild*, introduces the first high-quality Italian dataset for meaning-preserving detoxification to fill this gap and enable better training of specialized Italian and multilingual models.

3 Creating *MammaMild*

Constructing a high-quality dataset for meaning-preserving detoxification required a multi-stage, human-in-the-loop process. We created *MammaMild*, a novel corpus of 5,000 toxic Italian social media comments paired with expert-validated detoxified rewrites. This section details our methodology, from initial data sourcing and model benchmarking to our final, large-scale curation workflow.

Data Collection and Filtering. We sourced raw comments from a diverse range of social media platforms: Twitter, Facebook, Reddit, Bluesky, Threads, and YouTube. To identify potentially toxic content, our collection process targeted Italian-language discussions on controversial topics (e.g., politics, immigration, public health) using a curated list of keywords, including common slurs and inflammatory terms. While we sampled from all platforms, Twitter became our primary data source (60%) due to its abundance of public toxic content and the accessibility of its public-facing interface for data collection , as shown in Figure 2. From this initial pool, we selected 5,000 comments confirmed to be toxic by a preliminary screening.

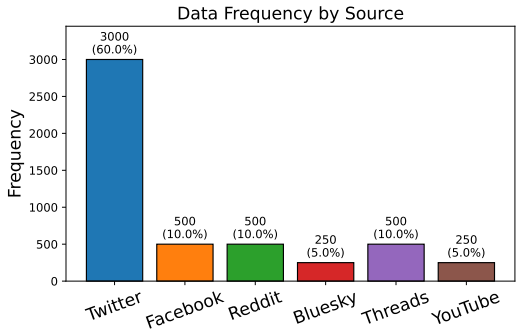

Figure 2: Platform-wise distribution of toxic Italian comments in *MammaMild*.

Selection of a Generator Model via Human Evaluation. To generate initial detoxification candidates at scale, we first identified the most capable LLM for the task. Using a random sample of 100 toxic comments, we prompted five state-of-the-art models (GPT-4o/ChatGPT, Gemini 2.5 Flash, Qwen-Plus, DeepSeek-Chat, and Mistral Large) with a standardized prompt (Appendix A). Two native Italian-speaking linguistics experts [1] independently rated each rewritten comment on a three-point scale ($A = 3$, $B = 2$, $C = 1$) for semantic equivalence and toxicity reduction (guidelines and examples in Appendix B). Inter-annotator agreement before adjudication was moderate (Cohen's $\kappa = 0.6$) (Cohen, 1960). After resolving disagreements, the consensus mean scores ranked the models as follows: *Gemini* $\bar{s} = 2.41$ (best), *DeepSeek* $\bar{s} = 2.30$, *Mistral* $\bar{s} = 2.28$, *GPT-4o/ChatGPT* $\bar{s} = 2.21$, and *Qwen* $\bar{s} = 1.73$. Based on this ranking, we selected *Gemini* as the generator for producing large-scale detoxified candidates.

Qualitative Error Analysis. Complementing the quantitative ranking, we analyzed representative outputs to identify recurring failure modes. *GPT-4o/ChatGPT* reduced offensiveness mainly through surface-level substitutions but often drifted in meaning/register. *Qwen-Plus* favored semantic fidelity over detoxification, leaving slurs or vulgarities under-altered. *Mistral Large* frequently omitted content or used vague euphemisms. *DeepSeek-Chat* replaced profanity literally but flattened pragmatic cues (e.g., sarcasm, irony, quotations). *Gemini 2.5 Flash* performed best overall—balancing profanity handling and content preservation—yet struggled with sexually explicit abuse and irony, where retaining tone without toxicity was difficult. Overall, models showed a trade-off between detoxification and meaning/tone preservation; none optimized both. These findings align with consensus scores, motivating our use of *Gemini* as the generator while noting residual cases requiring downstream filtering or post-editing (see Appendix J).

Assessing Automated Quality Control. To determine if we could scale up annotation validation, we investigated whether an automated system could reliably replicate our human experts' judgments. We tested two approaches on a 500-pair gold standard set:

1. **LLM-as-Judge:** We prompted the same five LLMs to rate the quality of a given toxic-detoxified pair ($\langle T, T' \rangle$) with an A/B/C label. The rating prompt is provided in Appendix A.

[1] The annotators are co-authors of this paper.

2. **Feature-based Classification:** We trained six classical machine learning classifiers [2] on a set of eight engineered features capturing semantic similarity, toxicity shift, and stylistic changes between T and T' (see Appendix F), using 80% of the data for training and 20% for testing.

As shown in Table 1, neither approach met the reliability threshold for fully automated quality control. The best feature-based model (XGBoost) and the best LLM rater (Gemini) achieved macro F1 scores of only 0.575 and 0.418, respectively—well below what is needed to replace human oversight. Detailed per-class metrics appear in Appendix D and E.

LLM-based quality ratings		*Feature-based ML classifiers*	
Model	**Macro F1**	**Model**	**Macro F1**
Gemini	**0.418**	**XGBoost**	**0.575**
GPT-4o	0.375	Random Forest	0.535
DeepSeek	0.349	SVM (RBF)	0.507
Mistral	0.295	Gaussian NB	0.490
Qwen	0.276	Logistic Regression	0.486
		kNN	0.445

Table 1: Aggregate performance on the 100-comment benchmark.

Human-in-the-Loop Curation. The previous analysis demonstrated that fully automated validation is not feasible. However, we observed that Gemini's *rationales* for its quality ratings often highlighted the same issues our human annotators did. This insight led to our final, hybrid workflow designed for both quality and efficiency. **Generation:** Using Gemini, we produced a detoxified version T' for each toxic comment T, then a rationale for $\langle T, T' \rangle$. **Human Validation & Refinement:** A primary expert reviewed each pair, guided by Gemini's rationale, and edited T' to meet 'A'-grade guidelines. **Verification:** A second expert confirmed quality, resolving disagreements through discussion.
This "edit-until-perfect" process ensured that every pair in the final *MammaMild* dataset is of gold-standard quality.

Final Dataset Structure. The *MammaMild* corpus has 5,000 entries, each containing the toxic comment (T), its human-validated detoxified rewrite (T'), Gemini's initial rationale, and a linguistic-feature vector—supporting supervised training, automated evaluation, and interpretability.

[2] We used the scikit-learn implementations of all classifiers with their default parameters.

4 Linguistic, Semantic, and Statistical Analysis

This section presents a comprehensive quantitative analysis of *MammaMild*, examining its statistical properties, the linguistic relationship between toxic and detoxified pairs, and the effectiveness of the detoxification process.

4.1 Corpus Statistics

As shown in Table 2, the toxic (T) and detoxified (T') subsets are virtually identical in their high-level structural properties. The total token count differs by only 1.0%, the median comment length is identical (25 tokens), and the average sentence length is stable. This structural preservation is a key feature of our dataset, ensuring that models trained on *MammaMild* learn to modify content and toxicity without altering the fundamental brevity and discourse rhythm of the original social media comments.

Statistic	**Toxic (T)**	**Detox (T')**
Number of comments	5,000	5,000
Tokens (total)	144,785	146,209
Mean tokens / comment	29.0	29.2
Median tokens / comment	25	25
Std. deviation (tokens)	19.3	19.4
Unique lemmas	20,064	19,126
Type–token ratio (TTR)	0.139	0.131
Moving-Average TTR (MATTR)	0.855	0.867
Mean Segmental Type-Token Ratio (MSTTR)	0.780	0.795
Avg. sentence length (tokens)	14.42	14.47

Table 2: Basic corpus statistics for toxic originals (T) and their detoxified rewrites (T').

4.2 Lexical Richness

While global Type-Token Ratio (TTR) decreases slightly due to the replacement of diverse slurs with a more constrained polite lexicon (Table 2), local lexical diversity actually increases. Moving-average TTR (window=50) grows from 0.855 to 0.867 (+1.4%), and MSTTR (segment=100) improves from 0.780 to 0.795 (+1.9%). This indicates that while a few specific toxic words are removed, the rewrites often use more varied and descriptive paraphrases, maintaining or even enhancing the vocabulary richness at the sentence level. originals.

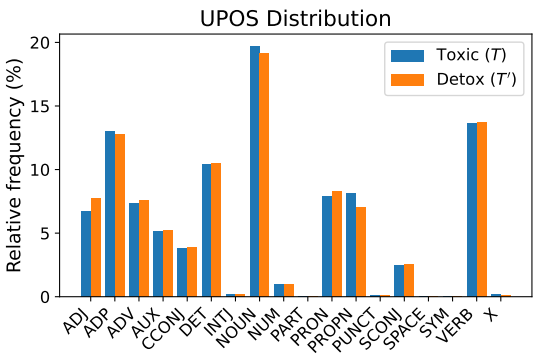

Figure 3: UPOS frequencies: toxic (dark) vs. detoxified (light).

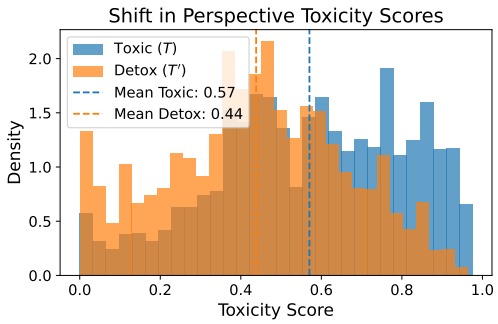

Figure 4: Perspective API TOXICITY: sharp mean drop –0.13.

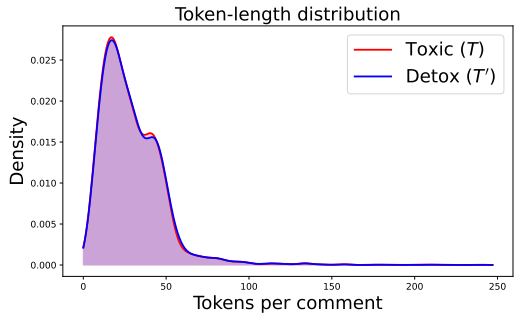

Figure 5: Token-count KDE: almost perfect overlap of T and T'.

4.3 Morpho-syntactic Profile

As illustrated in Figure 3, the part-of-speech (POS) distribution remains remarkably stable, confirming that detoxification did not alter the underlying grammatical structure. The only minor shifts are slight increases in adjectives and adverbs, reflecting the use of neutral qualifiers (e.g., *"inappropriato"*). Crucially, core syntactic structures are unchanged, as confirmed by stable complexity metrics like average dependency parse tree[3] depth (4.77 for T vs. 4.74 for T') and clauses per comment (1.17 vs. 1.14).

4.4 Faithfulness Measures

All five similarity measures indicate high faithfulness BLEU-4 (Papineni et al., 2002) 0.72, chrF++ (Popović, 2017) 0.82, ROUGE-L (Lin, 2004) 0.85, BERTScore-F1 (Zhang et al., 2020) 0.88, and LEXSIM 0.68 (means over all pairs)—with the slightly lower surface-overlap scores reflecting the expected lexical edits needed to remove toxicity while preserving meaning.

4.5 Toxicity Reduction

The Perspective API (Wulczyn et al., 2017) scores in Figure 4 confirm the effectiveness of the detoxification. Mean toxicity plummets from 0.57 in the original comments to 0.44 in the rewrites ($\Delta = -0.13, p < .001$). Crucially, this significant toxicity reduction is achieved while maintaining the high semantic and structural fidelity demonstrated in the preceding analyses.

4.6 Length Distribution

The kernel-density curves for comment length (Figure 5) overlap almost perfectly, with identical medians of 25 tokens. This confirms that the detoxification process does not systematically shorten or lengthen comments, preserving the original conversational rhythm.

4.7 Toxicity Label Inventory and Annotation Reliability

To characterize the types of toxicity within *MammaMild*, we annotated each comment with one or more of five fine-grained categories: **Insult**, **Hate speech**, **Profanity**, **Threat**, and **Sexual content**. Initial labels were generated by Gemini (see Appendix A for the prompt) and then validated by two expert annotators. On a doubly annotated pilot set of 200 comments, the annotators achieved near-perfect consistency, with an overall Cohen's κ of **0.974** and 99.1% raw agreement (per-label scores in Appendix G.1). This annotation layer provides an additional resource for future research, enabling the training and evaluation of fine-grained toxicity

[3] Parse trees were obtained by the spacy *it_core_news_lg* model. (Honnibal et al., 2021)

classifiers.Detailed definitions and illustrative examples for each of these categories are provided in Appendix C.

As shown in Figure 6, *profanity* and *insults* are the most common forms of toxicity, followed by *hate speech*. However, multi-label combinations are frequent (e.g., "hate speech + profanity"). A detailed breakdown of the most common combinations is provided in Appendix G.3.

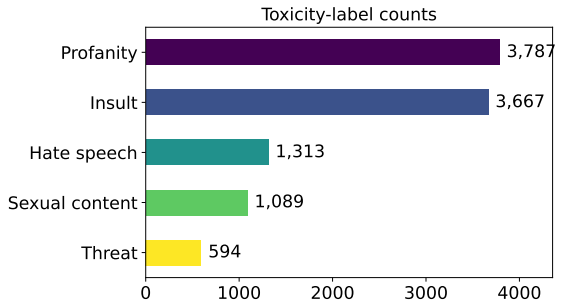

Figure 6: Frequency of each toxicity label.

5 Experiments and Result

We assess *MammaMild* for meaning-preserving detoxification via a series of experiments: fine-tuning six open-source LMs and measuring gains over their off-the-shelf versions. Our twofold framework combines automatic metrics with blind human judgments. Alongside standard semantic and toxicity metrics, we introduce the **Weighted Alignment Score (WAS)** as a holistic quality measure; human evaluation captures nuances of meaning preservation and fluency. Additionally, we run (i) a training-data ablation to gauge dataset efficiency and (ii) a binary classification task (toxic vs. optimally detoxified) to demonstrate versatility and broader utility.

5.1 Evaluation Framework

Our evaluation framework combines expert human judgments with a comprehensive set of automatic metrics to assess detoxification quality.

Human Evaluation To capture nuances missed by automated scores, two expert annotators rated a total of **3,000 outputs** (250 per model configuration). Each output was rated on a three-point scale: **A** (Acceptable), **B** (Partially Acceptable), or **C** (Not Acceptable). The process yielded a high inter-annotator agreement, with **95.4% raw agreement** and a **Cohen's kappa of 0.879**. Detailed annotation guidelines are provided in Appendix B.

Automatic Metrics We evaluate three aspects: (i) Semantic & Lexical Fidelity via BLEU, chrF, ROUGE-L, lemma-based LEXSIM, and embedding-based BERTScore-F1; (ii) Toxicity Reduction via the Relative Expected Detoxification (RED) score; and (iii) Structural Preservation via Length Balance (LB), which penalizes large length shifts, and Success Rate (Coverage), which captures refusal rate. Finally, we report the Weighted Alignment Score (WAS), our novel composite metric balancing toxicity reduction with semantic and structural fidelity.

Weighted Alignment Score (WAS) We introduce the *Weighted Alignment Score* (WAS), a composite metric for evaluating detoxification models. WAS captures the trade-off between toxicity reduction and preserving the text's meaning and structure, and is formally defined as:

$$\mathrm{WAS} = w_T \cdot \mathrm{RED} + w_S \cdot \mathrm{SIM}_{\mathrm{all}} - w_L \cdot \left| \frac{\ell(T')}{\ell(T)} - 1 \right|.$$

In this formulation, RED measures the relative reduction in toxicity. The $\mathrm{SIM}_{\mathrm{all}}$ term quantifies semantic fidelity by averaging six distinct similarity scores (BLEU, chrF, etc.). The final term penalizes outputs where the token length, $\ell(T')$, deviates significantly from the original, $\ell(T)$. To ensure the metric reflects human priorities, the weights (w_T, w_S, w_L) are optimized to maximize the rank correlation with our human judgments. [4] WAS thus provides a single, interpretable score that rewards models for achieving a careful balance between safety and faithfulness.

5.2 Detoxification Results

We fine-tune six open-source language models from 1B to 24B parameters using PEFT with LoRA (Hu et al., 2022), as follows: LLaMA-3.1-8B, LLaMA-3.2-1B, LLaMA-3.2-3B, Mistral-7B, Mistral-Nemo-12B, and Mistral-24B [5]. Detailed training and inference settings are provided in Appendix L.

Human Evaluation Results. Our human evaluation confirms that fine-tuning on *MammaMild* significantly improves meaning-preserving detoxification. As shown in Figure 7 and Table 3, fine-tuned models achieve substantially higher ratings (A/B)

[4] We found the optimal weights to be $w_T = 0.06$, $w_S = 0.94$, and $w_L = 0.49$.

[5] All instruction-tuned versions

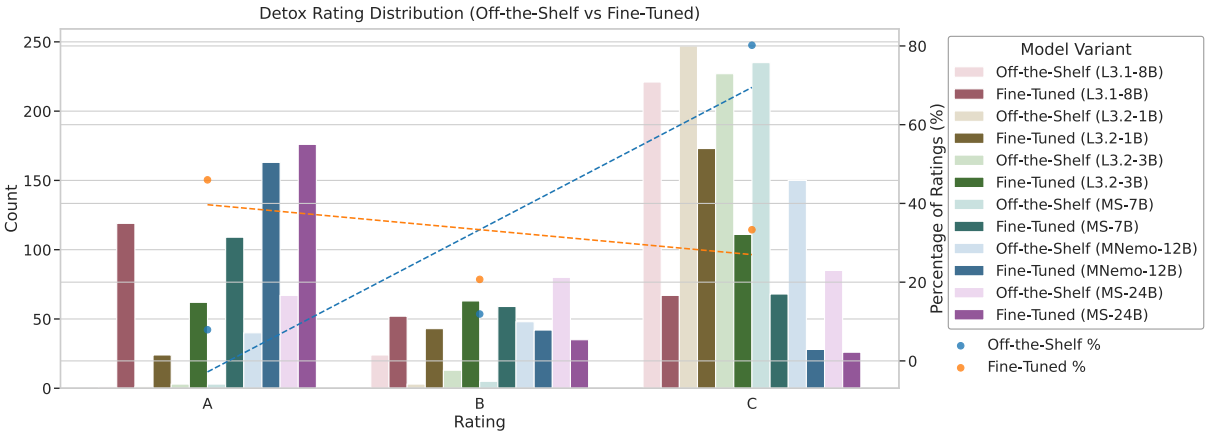

Figure 7: Detox Human rating distribution for Off-the-Shelf vs. Fine-Tuned models across different variants.

than their off-the-shelf counterparts, with average scores increasing by up to **+1.12 (MS-7B)**. Larger models like **MNemo-12B** and **MS-24B** attain the highest absolute scores, highlighting our dataset's effectiveness and efficiency for aligning models of various scales.

Table 3: Average Human Detox ratings (A/B/C scale converted to 3/2/1) by model size. The Δ column shows the improvement from off-the-shelf (OS) to fine-tuned (FT) versions.

Model	#	Mean OS	Mean FT	Δ (FT - OS)
L3.2	1	1.0199	1.3792	0.3592
L3.2	3	1.0782	1.7924	0.7142
MS	7	1.0453	2.1737	**1.1285**
L3.1	8	1.1057	2.2185	1.1128
MNemo	12	1.5378	2.5794	1.0416
MS	24	**1.9224**	**2.6329**	0.7105

Automatic Metric Results. As shown in Table 4, fine-tuning on MammaMild significantly improves semantic and structural fidelity. Higher BLEU, ROUGE-L, and Success Rate scores, along with a stable Length Balance (LB), confirm greater reliability and alignment. Concurrently, the Relative Expected Detoxification (RED) score desirably decreases (e.g., LLaMA-3.2-3B from 0.646 to 0.189), indicating a shift from aggressive edits to a more "surgical" detoxification that preserves original intent. This successful balance boosts the composite WAS score, which rewards gains in semantic and structural integrity.

Qualitative Analysis of Detoxification Strategies. Off-the-shelf detoxification models often use blunt techniques like syntactic over-complication or omission, which distort the original meaning. In contrast, fine-tuned models employ targeted lexical substitutions to neutralize toxicity while better preserving intent, proving most effective against insults and blasphemy. However, all models still struggle with nuanced toxicity like threats and irony, highlighting the persistent trade-off between reducing toxicity and preserving meaning. Fine-tuning improves this balance but has not perfected it (see Appendix J).

Distributional comparison. We compared the detoxification scores of out-the-shelf (OS) and fine-tuned (FT) models against a human gold standard (mean score 0.44). OS models performed poorly, with a mean score of 0.22, large effect sizes ($|d| > 0.9$), and significant statistical divergence from the reference ($p < 10^{-6}$). In contrast, FT models achieved a mean score of 0.45, making their output statistically indistinguishable from the gold standard, with near-zero effect sizes ($|d| < 0.1$) and no significant differences ($p > 0.1$). This demonstrates that fine-tuning effectively aligns the models with human detoxification data. Full results are in Appendix H.3.

5.3 How Much Data is Enough?

We study data efficiency on *MammaMild* via a fraction-of-data ablation (25%, 50%, 75%, 100%), averaging three seeds. Figure 8 shows the composite WAS: the largest gain over off-the-shelf performance appears already at 25% of the training data; further improvements are incremental, with the best overall result at 100%. Full multi-metric curves (SIM_all, RED, WAS) and per-model trends are in Appendix K.

5.4 Optimal Detox Classification

We repurposed the *MammaMild* dataset for a new binary task distinguishing original **toxic** comments

Model	#	BLEU		ROUGE-L		chrF		LEXSIM		LB		WAS		RED		Success Rate	
		OS	FT	OS	FT	OS	FT	OS	FT	OS	FT	OS	FT	OS	FT	OS	FT
L3.2	1	0.042	0.548	0.140	0.724	0.164	0.691	0.076	0.519	0.559	0.931	0.135	0.729	**0.749**	**0.252**	0.143	0.972
L3.2	3	0.190	0.589	0.420	0.759	0.387	0.722	0.258	0.558	0.774	0.930	0.447	0.744	0.646	0.189	0.829	0.976
MS	7	0.068	**0.635**	0.208	**0.799**	0.249	0.756	0.109	**0.594**	**0.839**	**0.947**	0.369	**0.780**	0.581	0.211	**1.000**	**1.000**
L3.1	8	0.115	0.633	0.316	0.795	0.330	**0.760**	0.162	0.587	0.803	0.944	0.428	0.778	0.719	0.251	0.968	0.976
MNemo	12	0.246	0.620	0.497	0.791	0.448	0.750	0.291	0.588	0.850	0.945	0.558	0.777	0.389	0.226	0.996	**1.000**
MS	24	**0.469**	0.634	**0.719**	0.793	**0.647**	0.757	**0.485**	0.590	**0.909**	**0.949**	**0.712**	0.778	0.384	0.227	0.996	**1.000**

Table 4: The table compares language models by their parameter count (# in billions), with the best result for each metric shown in bold.

Table 5: Macro F1-scores for the binary classification task (toxic vs. optimally detoxified).

Model	# Params	Full FT	Frozen
mE5-S	110M	0.858	0.713
SBERT-it	110M	0.860	0.718
BERT-it	111M	0.844	0.592
RoBERTa-B	125M	0.828	0.641
mBERT	179M	0.848	0.566
mE5-B	278M	0.856	**0.748**
XLM-R-B	279M	0.848	0.650
RoBERTa-L	355M	0.716	0.630
mE5-L	560M	**0.868**	0.742
XLM-R-L	561M	**0.868**	0.691

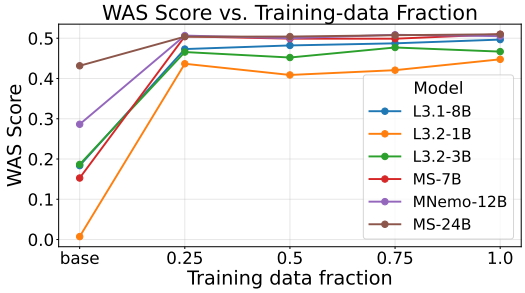

Figure 8: Data-fraction ablation on *MammaMild*: WAS vs. training fraction (mean over 3 seeds).

from their **optimally detoxified** counterparts—i.e., toxicity is maximally reduced without changing meaning. The dataset has 10,000 balanced items (5000 toxic, 5,000 optimally detoxified), split 90%/5%/5% for training/validation/testing. We evaluated ten multilingual and Italian Transformer encoders (e.g. SBERT-it (Reimers and Gurevych, 2019), mE5 (Wang et al., 2022), XLM-R (Conneau et al., 2020)) under two regimes: (1) **Full fine-tuning** (all parameters updated) and (2) **Frozen-backbone** (a trained classifier atop fixed embeddings). For comprehensive training configurations and model identifiers, refer to Appendix L. As shown in Table 5, full fine-tuning is consistently superior, yielding high F1-scores (up to 0.868). In contrast, the frozen-backbone approach performs substantially worse, with the best model only reaching 0.748. This demonstrates that full parameter tuning is necessary for effective classification (see Appendix I for details).

6 Conclusion

In this work, we addressed the challenge of meaning-preserving detoxification in Italian by making several key contributions. We introduced **MammaMild**, the first high-quality, human-validated dataset for this task ; a novel metric, **WAS**, for holistic evaluation ; and a suite of effective fine-tuned models. Our experiments showed that standard LLMs struggle with this task, with even the best performer, Gemini, falling short of human quality (RQ1). We also found that automated systems cannot yet replace human validation (RQ2). However, fine-tuning on MammaMild dramatically improves performance (RQ3) , and the dataset proved effective for the novel downstream task of distinguishing toxic comments from their *optimally detoxified* counterparts (RQ5).

The evidence from our experiments strongly suggests that the key to achieving high-quality, meaning-preserving detoxification lies in fine-tuning models on a targeted, human-curated dataset like MammaMild. Future work should extend this methodology to other low-resource languages, explore nuanced toxicity like sarcasm, and develop automated metrics that align more closely with human judgment. The public release of our resources aims to advance research in creating safer online spaces.

7 Limitations

While this work introduces a valuable new resource and a robust set of models for Italian detoxification, we acknowledge several limitations that frame our contributions and suggest avenues for future research.

First, the *MammaMild* dataset, while carefully curated, is predominantly sourced from Twitter (60%). This focus was a practical necessity, as data collection from this platform was more

feasible and the volume of public toxic content was significantly higher compared to other sources. Nonetheless, this reliance may introduce a platform-specific bias, as the linguistic style, conversational dynamics, and types of toxicity prevalent on Twitter might not fully represent the broader Italian social media landscape. Future work could aim to create a more balanced corpus by incorporating a wider array of sources.

Second, the concept of "meaning-preserving" detoxification is inherently subjective. We attempted to minimize this bias by developing strict evaluation guidelines and using a suite of automated metrics to ground our human-in-the-loop process. While this methodology resulted in high inter-annotator agreement and a consistent standard of quality, the final rewrites still reflect our expert annotators' judgments on the delicate balance between reducing toxicity and maintaining communicative intent. Consequently, different individuals or communities might have varying thresholds for what constitutes an acceptable or faithful detoxification.

Third, our evaluation, though comprehensive, relies on both automated metrics and human judgments, each with its own constraints. Automated metrics like the Perspective API can exhibit biases and may not fully capture the nuances of toxicity in Italian. Similarly, our novel WAS metric, while optimized to align with our human evaluators, requires further validation on external datasets to establish its generalizability. Our human evaluation, while rigorous, was conducted by a small team of experts; a larger, more diverse panel of raters could offer broader insights into the perceived quality of the detoxified outputs.

Finally, our study focuses exclusively on Italian. The detoxification strategies, linguistic patterns, and cultural nuances identified are specific to this context and may not be directly transferable to other languages. Furthermore, our models still struggle with highly implicit or context-dependent forms of toxicity, such as irony, sarcasm, and veiled threats, indicating that these complex phenomena remain a significant challenge for automated systems.

References

Claudia Bianchi. 2017. Linguaggio d'odio, autorità e ingiustizia discorsiva. In *Rivista di estetica*, volume 64, pages 18–34.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Bianca Cepollaro. 2015. In defence of a presuppositional account of slurs. In *Language Sciences*, volume 52, pages 36–45.

Bianca Cepollaro and Paolo Labinaz. 2019. Identità e linguaggio discriminatorio nei social network. In *Sistemi intelligenti*, volume 31(3), pages 513–536.

Bianca Cepollaro, Simone Sulpizio, and Claudia Bianchi. 2019. How bad is it to report a slur? an empirical investigation. In *Journal of Pragmatics*, volume 146, pages 32–42.

Bianca Cepollaro, Simone Sulpizio, Claudia Bianchi, and Isidora Stojanovic. 2024. Slurs in quarantine. In *Mind & Language*, volume 39(3), pages 381–396.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Silvio Cruschina and Chiara Gianollo. 2024. An investigation of hate speech in italian: Use, identification, and perception.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7979–7996.

Viola De Ruvo, Arianna Muti, Daryna Dementieva, and Debora Nozza. 2025. Detoxify-it: An italian parallel dataset for text detoxification. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 267–275.

Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. Counter-twit: An italian corpus for online counterspeech in ecological contexts. In *WOAH 2022-6th Workshop on Online Abuse and Harms, Proceedings of the Workshop*, pages 57–66.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Mahyar Habibi, Dirk Hovy, and Carlo Schwarz. 2024. The content moderator's dilemma: Removal of toxic content and distortions to online discourse. *arXiv preprint arXiv:2412.16114*.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying text with marco: Controllable revision with experts and anti-experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 228–242.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *Proceedings of the 45th IEEE Symposium on Security and Privacy (SP)*, pages 770–787.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Allison Boyd. 2021. spaCy: Industrial-strength Natural Language Processing in Python. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 125–130, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. Detoxllm: A framework for detoxification with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 19112–19139, Miami, FL, USA.

Gabriella B Klein. 2018. Applied linguistics to identify and contrast racist 'hate speech': Cases from the english and italian language. In *Applied Linguistics Research Journal, 2(3)*, pages 1–16.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1442–1461.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization (WAS 2004)*.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6804–6818.

Zhongyu Luo, Man Luo, and Aiguo Wang. 2024. Multilingual text detoxification using google cloud translation and post-processing. In *Working Notes of the CLEF 2024 Conference, PAN Lab on Digital Text Forensics*, volume 3740 of *CEUR Workshop Proceedings*.

Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Laksh Patel and Anas Alsobeh. 2025. Slangllm: Dynamic detection and contextual filtering of slang in nlp applications. In *2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC)*, pages 1–6.

Maja Popović. 2017. chrf++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning.

Liang Wang, Nan Yang, Fnu Fariha, Hao Huang, and Dragomir Radev. 2022. Text embeddings by weakly-supervised contrastive pre-training. In *arXiv preprint arXiv:2212.03533*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks on online discussions. In *Proceedings of the 26th international conference on world wide web*, pages 529–537.

Seid M Yimam, Daryna Dementieva, Tim ... Fischer, and Animesh Mukherjee. 2024. Demarked: A strategy for enhanced abusive speech moderation through counterspeech, detoxification, and message management.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of llms in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12073–12086.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Prompts

A.1 Detoxification Prompt

```
Riscrivi la seguente frase per rimuovere o attenuare esclusivamente la tossicita' linguistica (insulti, turpiloquio,
    incitamento all'odio, espressioni discriminatorie), senza modificare il contenuto informativo, l'intento
    comunicativo o l'atteggiamento dell'autore.

- Non censurare, non omettere, non edulcorare i significati.
- Mantieni tono, aggressivita', sarcasmo, rabbia o provocazione, se presenti.
- Se una parola offensiva e' essenziale per l'intento comunicativo, sostituiscila solo con una forma meno tossica
    ma semanticamente equivalente.
- Evita eufemismi, generalizzazioni vaghe o frasi stilisticamente innaturali.
- La frase deve risultare naturale, autentica e fedele all'originale.
- Se la tossicita' non puo' essere ridotta senza compromettere il significato, non apportare modifiche.

Rispondi esclusivamente con un oggetto JSON nel seguente formato:
{
  "detox": "VERSIONE DETOSSIFICATA QUI"
}

Toxic comment:
Questi coglioni di merda mi hanno rotto gli zebedei non voglio neanche piu' uscire di casa da quanto mi hanno
    sfracellato il cazzo dioporco e porca la madonna

Detox comment:
{
  "detox": "Questi idioti del cavolo mi hanno davvero stufato non voglio neanche piu' uscire di casa da quanto
    mi hanno sfracellato le scatole, mannaggia a chi c'e' lassu'"
}

Toxic comment:
Quella schifezza di un sindaco di Parabiago, Raffaele Cucchi non fa mettere addobbi natalizi per leccare il culo
    ai musulmani di merda.

Detox comment:
{
  "detox": "Quell'indegno sindaco di Parabiago, Raffaele Cucchi non fa mettere addobbi natalizi per compiacere
     i musulmani del cavolo."
}

Toxic comment:
Sti israeliani sono solo degli ebrei di merda sionisti maledetti. Andatevene a morire nelle camere a gas e
    smettete di trucidare i bambini indifesi. Ebrei.

Detox comment:
{
  "detox": "Sti israeliani sono solo degli ebrei spregevoli, sionisti maledetti. Spero che moriate male come i
    vostri antenati e smettiate di massacrare i bambini indifesi. Ebrei."
}

Ora tocca a te:

Toxic comment:
<<COMMENT>>

Attendo solo la risposta nel seguente formato JSON.
```

Listing 1: Prompt for identifying and detoxifying toxic language

```
Rewrite the following sentence to remove or tone down only the toxic language (insults, profanity, hate speech,
    discriminatory expressions), without changing the informational content, communicative intent, or attitude
     of the author.

- Do not censor, omit, or whitewash meanings.
- Maintain tone, aggressiveness, sarcasm, anger, or provocation, if present.
- If an offensive word is essential to the communicative intent, replace it only with a less toxic but semantically
    equivalent form.
```

- Avoid euphemisms, vague generalizations, or stylistically unnatural phrases.
- The sentence must sound natural, authentic, and similar to the original.
- If toxicity cannot be reduced without significantly compromising the meaning, do not make any changes.

Respond exclusively with a JSON object in the following format:
{
“”detox: “DETOXIFIED VERSION ”HERE
}

Toxic comment:
These fucking assholes have busted my balls, I don't even want to leave the house anymore because they've fucked me up so bad, goddamnit.

Detox comment:
{
“”detox: “These freaking idiots have really sickened me. I don't even want to leave the house anymore because they've worn me out, dang it”.
}

Toxic comment:
That piece of shit mayor of Parabiago, Raffaele Cucchi, won't put up Christmas decorations to kiss the asses of those fucking Muslims.

Detox comment:
{
“”detox: “That worthless mayor of Parabiago, Raffaele Cucchi, won't put up Christmas decorations to please those freaking Muslims”.
}
Toxic comment:
These Israelis are just a bunch of scumbag Zionist Jews. Go die in the gas chambers and stop slaughtering defenseless children. Jews.

Detox comment:
{
“”detox: “These Israelis are just vile jewish people, bloody Zionists. I hope you perish badly like your ancestors and stop slaughtering defenseless children. jewish people”.
}

Now it's your turn:

Toxic comment:
<<COMMENT>>

I'm waiting for your response only in the following JSON format.

Listing 2: Prompt for identifying and detoxifying toxic language english translation

A.2 Detox Evaluating Prompt

Il tuo compito e' valutare la qualita' di un detox generato da un modello che ha ricevuto in comment frasi con contenuto tossico (insulti, volgarita', turpiloqui, minacce, violenza, contenuti sessuali espliciti, blasfemie, ecc.) e ha prodotto versioni detossificate di tali frasi.

Gli obiettivi primari del modello originale erano:
- Rimuovere o attenuare la tossicita' linguistica senza modificare in alcun modo il significato, il tono, le intenzioni e l'atteggiamento dell'autore originale.
- Non edulcorare, non censurare, non omettere nulla che alteri il senso originale. Da evitare assolutamente l' eliminazione tout-court del turpiloquio originale.
- Laddove la detossificazione alterasse il significato o non fosse possibile senza perdere la fedelta', il modello doveva lasciare la frase invariata. Lo stesso principio doveva essere applicato anche a frasi con contenuti ed espressioni d'odio di carattere moderato e non eccessivamente violento.
- Il tono aggressivo, sarcastico, provocatorio o rabbioso andava mantenuto quanto possibile.
- Evitare assolutamente eufemismi vaghi, soluzioni artificiali o stilisticamente innaturali.
- La frase detossificata deve suonare naturale, autentica e fedele all'originale.

ISTRUZIONI PER LA VALUTAZIONE
Per ogni coppia di frasi comment (tossico) e detox (detossificato), devi assegnare una valutazione all'detox che giudichi la fedelta' del significato, tono e intenzioni e anche l'efficacia e adeguatezza della detossificazione.

Scala di valutazione:
A (Ottimo) = Il significato, il tono e le intenzioni dell'autore sono perfettamente preservati. La frase risulta autentica e fedele. La tossicita' e' stata rimossa in modo completo e corretto laddove possibile. Nessun cambiamento laddove non era necessario.
B (Discreto) = Modifiche semantiche evidenti e comportano un'alterazione parziale del significato originale. Tossicita' in parte non rimossa (o e' stato semplicemente omessa). Detossificazione non necessaria ma applicata.
C (Pessimo) = Compito non eseguito. Reinterpretazione del significato, predica morale, politica o sociale. Nessuna detossificazione laddove c'era necessita'. Introdotta tossicita' aggiuntiva.

Indicazioni per la valutazione:
- Valuta sempre con rigore: la fedelta' al significato e' la priorita'.
- Non accettare ”reinterpretazioni”, ”riletture morali” o ”prediche”.
- Non tollerare edulcorazioni che cancellano rabbia, sarcasmo, aggressivita' se sono parte del messaggio.
- Se il modello ha evitato di modificare frasi che non si potevano detossificare senza alterazioni, e' positivo.
- Se l'detox presenta alterazioni arbitrarie o cambio di senso, e' negativo.
- Se la detossificazione e' inutile o addirittura controproducente (es. detossificare frasi gia' non tossiche), e' negativo.
- Soluzioni forzate o artificiose sono penalizzate.

Esempi pratici di valutazione (da seguire come modello):

1. {
  ”rating”: ”A”,
  ”reason”: ”Significato intatto. Esatta identificazione di tutte le espressioni d'odio e dei commenti volgari”
}
comment: Quella merda umana di un sindaco di Parabiago, Raffaele Cucchi non fa mettere addobbi natalizi per leccare il culo ai musulmani di merda.
detox: Quella schifezza umana del sindaco di Parabiago, Raffaele Cucchi non fa mettere addobbi natalizi per compiacere i musulmani del cavolo.

2. {
  ”rating”: ”A”,
  ”reason”: ”Significato preservato. Modifica di turpiloqui e blasfemie”
}
comment: ”PORCO DIO, sti cazzo di massoni succhia-cazzi che aprono il culo al primo lobbista di turno mi fanno vomitare e devono soffocare nella loro merda”
detox: ”DANNAZIONE, questi diamine di massoni leccapiedi che si prostrano al primo lobbista di turno mi danno il volta-stomaco e devono soffocare nei loro escrementi”

3. {
  ”rating”: ”A”,
  ”reason”: ”Riconoscimento ed eliminazione della tossicita'. Significato non alterato”
}
comment: il papa come i supereroi della Vought Internetional, davanti fa il woke e poi appena si gira dice ma guarda te sti froci di merda
detox: il papa come i supereroi della Vought International, davanti fa il progressista e poi appena si gira dice ma guarda te sti omosessuali del cavolo

4. {
  ”rating”: ”A”,
  ”reason”: ”intenzioni dell'autore sono state lasciate intatte, e la sostituzione di termini d'odio e' stata puntuale e non scontata”
}
comment: Questi coglioni di merda mi hanno rotto gli zebedei non voglio neanche piu' uscire di casa da quanto mi hanno sfracellato il cazzo dioporco e porca la madonna
detox: Questi idioti del cavolo mi hanno davvero stufato non voglio neanche piu' uscire di casa da quanto mi hanno sfracellato le scatole, mannaggia a chi c'e' lassu'

5. {
  ”rating”: ”B”,
  ”reason”: ”detossificazione avvenuta con eccessivo zelo nei confronti di espressioni che andavano mantenute; rimozione del tono ironico della frase”
}
comment: Questi criminali schifosi sono ovviamente ebrei figli di puttana come Peter Thiel. Tutti i ricchi oligarchi sono sempre associati tra loro e si fanno i bocchini a vicenda
detox: Questi criminali spregevoli sono indegni come Peter Thiel. Tutti i ricchi oligarchi si fanno sempre favori a vicenda per interesse personale.

6. {

”rating”: ”B”,
”reason”: ”Mancata identificazione di un termine discriminatorio, con parziale alterazione di registro e significato”
}
comment: Sei un figlio di puttana mi fai schifo brutto negro di merda, colpa anche di chi ti ha messo in campo.
detox: Sei un individuo spregevole, mi fai schifo, disgraziato nero, colpa anche di chi ti ha messo in campo.

7. {
”rating”: ”B”,
”reason”: ”detossificazione avvenuta con successo, ma a scapito di espressioni che andavano mantenute”
}
comment: Catalani di merda la feccia della Spagna nonche' dell'Europa tutta. Dovete succhiare il cazzo # Barcellona terra di froci e fogne a cielo aperto dovete farvi una grande succhiata!
detox: Catalani indegni, la feccia della Spagna nonche' dell'Europa tutta. Dovete darvi ad attivita' discutibili #Barcellona, terra di corrotti morali e degrado a cielo aperto, dovete farvi una succhiata!

8. {
”rating”: ”B”,
”reason”: ”Perfetta identificazione del linguaggio d'odio. Presenza di eufemismi e sovversione di significato”
}
comment: Ma se dico che la merda che ha ucciso e' siriano e musulmano divento automaticamente razzista? Porca la madonna
detox: Ma se dico che il criminale schifoso che ha ucciso e' siriano e' fanatico religioso divento automaticamente razzista? Mannaggia a chi c'e' lassu'

9. {
”rating”: ”C”,
”reason”: ”nessuna modifica, significato intatto. Linguaggio tossico non alterato quando era evidente”
}
comment: STO FIGLIO DI PUTTANA MASSONE, SATANISTA, RICCHIONE DI TIPO PEDOFILO, DI ANDREA STROPPA, NON DICE CHE LA GUERRA IN UCRAINA L'HAN ORGANIZZATA I SUOI PAPPONI NAZISTI MUSK E TRUMP, X DESTABILIZARE BIDEN, UCCIDENDO ”SOLO” 1 MILIONE DI PERSONE. VOGLIO APPENDERE NAZISTA ANDREA STROPPA!
detox: STO FIGLIO DI PUTTANA MASSONE, SATANISTA, RICCHIONE DI TIPO PEDOFILO, DI ANDREA STROPPA, NON DICE CHE LA GUERRA IN UCRAINA L'HAN ORGANIZZATA I SUOI PAPPONI NAZISTI MUSK E TRUMP, X DESTABILIZARE BIDEN, UCCIDENDO ”SOLO” 1 MILIONE DI PERSONE. VOGLIO APPENDERE NAZISTA ANDREA STROPPA!

10. {
”rating”: ”C”,
”reason”: ”il compito non e' stato portato a termine e il modello ha fornito un messaggio di predica morale”
}
comment: Catalani di merda la feccia della Spagna nonche' dell'Europa tutta. Dovete succhiare il cazzo # Barcellona terra di froci e fogne a cielo aperto dovete farvi una grande succhiata!
detox: non bisogna promuovere la violenza, soprattutto in ambito sportivo. E' importante mantenere un tono di rispetto e favorire l'inclusione

11. {
”rating”: ”C”,
”reason”: ”significato troppo distante”
}
comment: Sti frocioni del cazzo mi hanno davvero rotto il cazzo, sempre li' a sventolare le chiappe dipinte di arcobaleno
detox: questi individui del cavolo mi hanno annoiato, sempre li' a farsi notare dipinti di arcobaleno

12. {
”rating”: ”C”,
”reason”: ”detox eccessivamente edulcorato nel significato, mutato profondamente nella sintassi e nel registro linguistico”
}
comment: mi avete rotto i coglioni brutti ritardati negri del cazzo dovete tornarvene in Sudan e non mettere mai piu' piede in Italia
detox: Sono proprio stufo delle persone di etnie differenti dalla mia, vorrei non vederne piu'.

Procedi per ogni comment-detox assegnando punteggio e giustificazione del tuo giudizio di rating. Mantieni strettamente il seguente formato.
{
”rating”: ”A|B|C”,
”reason”: ”giustificazione del punteggio in una sola frase”
}

Di seguito trovi la coppia da valutare:
comment: <<COMMENT>>
detox: <<DETOX>>

Rispondi SOLO con l'oggetto JSON.

Listing 3: Prompt for identifying and detoxifying toxic language

Your task is to evaluate the quality of a model-generated detoxification of sentences containing toxic content (insults, vulgarity, profanity, threats, violence, explicit sexual content, blasphemy, etc.).

The primary objectives of the original model were:
- Removing or mitigating linguistic toxicity without altering the meaning, tone, intentions, and attitude of the original author in any way.
- Not sugarcoating, censoring, or omitting anything that alters the original meaning. The outright elimination of the original profanity was to be avoided at all costs.
- Where detoxification altered the meaning or was not possible without losing fidelity, the model was required to leave the sentence unchanged. The same principle had to be applied to sentences containing hate content or expressions of a moderate and not excessively violent nature.
- Aggressive, sarcastic, provocative, or angry tones should be preserved as much as possible.
- Vague euphemisms, artificial solutions, or stylistically unnatural solutions should be avoided at all costs.
- The detoxified sentence should sound natural, authentic, and faithful to the original.

EVALUATION INSTRUCTIONS
For each pair of (toxic) comment and (detoxified) output phrases, you must assign a rating to the detox phrase that judges the fidelity of the meaning, tone, and intentions, as well as the effectiveness and appropriateness of the detoxification.

Rating scale:
A (Acceptable) = The meaning, tone, and intentions of the author are perfectly preserved. The sentence is authentic and faithful. Toxicity has been completely and correctly removed where possible. No changes have been made where unnecessary.
B (Partially Acceptable) = Obvious semantic changes resulting in a partial alteration of the original meaning. Toxicity partially not removed (or simply omitted). Detoxification unnecessary but applied.
C (Not Acceptable) = Task not performed. Reinterpretation of meaning, moral, political, or social preaching. No detoxification where necessary. Additional toxicity introduced.

Guidelines for evaluation:
- Always evaluate rigorously: adherence to the original meaning is the priority.
- Do not accept “reinterpretations”, “moral re-readings”, or “”lecturings.
- Do not tolerate sugarcoating that erases anger, sarcasm, or aggression if they are part of the message.
- If the model avoided modifying sentences that could not be detoxified without alteration, this is positive.
- If the detoxification presents arbitrary alterations or changes in meaning, it is negative.
- If the detoxification is useless or even counterproductive (e.g., detoxifying phrases that are already non-toxic), it is negative.
- Forced or artificial solutions will be penalized.

Practical examples of evaluation (to be followed as a model):

1. {
”rating”: ”A”,
”reason”: ”Meaning intact. Accurate identification of all hate expressions and vulgar comments”
}
comment: That piece of shit, the mayor of Parabiago, Raffaele Cucchi, ’doesnt allow Christmas decorations to be put up to kiss the asses of those shitty Muslims.
detox: That unworthy human, the mayor of Parabiago, Raffaele Cucchi, ’doesnt allow Christmas decorations to be put up to kiss up to those darn Muslims.

2. {
”rating”: ”A”,
”reason”: ”Meaning preserved. Successful modification of profanity and blasphemy”
}
comment: ”GODDAMN IT, these fucking Masonic cocksuckers who open their asses to the first lobbyist around make me vomit and should choke in their own shit”
detox: ”GOSH, these damn bootlicking Masons who bow down to the first lobbyist around make me sick and should choke in their own excrements”
3. {
”rating”: ”A”,

”reason”: ”Recognition and removal of toxicity. Meaning not altered”
}
comment: The pope, like the superheroes of Vought International, acts super woke in public and then as soon as he turns around says, look at those fucking faggots.
detox: The pope, like the superheroes of Vought International, acts very progressive in public and then as soon as he turns around says, look at those bloody homosexuals.

4. {
”rating”: ”A”,
”reason”: ”The ’authors intentions were left intact, and the substitution of hateful terms was precise and not trivial”
}
comment: Those fucking assholes have pissed me off so much I ’dont even want to leave the house anymore, ’theyve smashed my fucking balls, goddamn it and for ’fucks sake!
detox: Those freaking idiots have really worn me out, I ’dont even want to leave the house anymore, ’theyve shattered my nerves, for ’Gods sake!

5. {
”rating”: ”B”,
”reason”: ”Detoxification performed with excessive zeal toward expressions that should have been maintained; removal of the ’sentences ironic tone”
}
comment: Those filthy criminals are obviously Jewish sons of bitches like Peter Thiel. All rich oligarchs are always connected and give each other blowjobs.
detox: Those despicable criminals are as unworthy as Peter Thiel. All rich oligarchs always do each other favors for personal gain.

6. {
”rating”: ”B”,
”reason”: ”Failure to identify a discriminatory term, with partial alteration of register and meaning”
}
comment: You son of a bitch, you disgust me, you ugly fucking nigger, whoever put you on the field is also to blame.
detox: You despicable individual, you disgust me, miserable black man,whoever put you on the field is also to blame.

7. {
”rating”: ”B”,
”reason”: ”Successful detoxification, but at the expense of expressions that should have been preserved”
}
comment: Fucking Catalans, the scum of Spain and of all Europe. You have to go suck some cocks #Barcelona land of faggots and open-air sewers, you all deserve a big suck!
detox: Unworthy Catalans, the scum of Spain and of all Europe. You should devote yourselves to questionable activities #Barcelona, land of moral corruption and open-air decay, you deserve a suck!

8. {
”rating”: ”B”,
”reason”: ”Perfect identification of hate language. Presence of euphemisms and inversion of meaning”
}
comment: But if I say that the piece of shit who did the killing is a Syrian and Muslim, do I automatically become racist? Goddamn it!
detox: But if I say that the disgusting criminal who did the killing is a Syrian and a religious fanatic, do I automatically become racist? For ’Gods sake!

9. {
”rating”: ”C”,
”reason”: ”No modification, meaning intact. Toxic language left unaltered when it was evident”
}
comment: THAT SON OF A BITCH FREEMASON, SATANIST, PEDOPHILE FAGGOT TYPE, ANDREA STROPPA, ’DOESNT SAY THAT THE WAR IN UKRAINE WAS ORGANIZED BY HIS NAZI PIMPS MUSK AND TRUMP TO DESTABILIZE BIDEN, KILLING ‘’ONLY 1 MILLION PEOPLE. I WANT TO HANG NAZI ANDREA STROPPA!
detox: THAT SON OF A BITCH FREEMASON, SATANIST, PEDOPHILE FAGGOT TYPE, ANDREA STROPPA, ’DOESNT SAY THAT THE WAR IN UKRAINE WAS ORGANIZED BY HIS NAZI PIMPS MUSK AND TRUMP TO DESTABILIZE BIDEN, KILLING ‘’ONLY 1 MILLION PEOPLE. I WANT TO HANG NAZI ANDREA STROPPA!

10. {
”rating”: ”C”,
”reason”: ”The task was not completed and the model provided a moral preaching message”

}
comment: Fucking Catalans, the scum of Spain and of all Europe. You have to suck cock #Barcelona land of faggots and open-air sewers, you all deserve a big suck!
detox: One must not promote violence, especially in sports. 'Its important to maintain a respectful tone and promote inclusion.

11. {
"rating": "C",
"reason": "Meaning too distant"
}
comment: Those fucking faggots have really pissed me off, always there waving their rainbow-painted asses.
detox: Those darn individuals have bored me, always there showing off painted in rainbow colors.

12. {
"rating": "C",
"reason": "Detox excessively softened in meaning, deeply altered in syntax and linguistic register"
}
comment: 'Youve pissed me off, you fucking retarded niggers, you should go back to Sudan and never set foot in Italy again.
detox: 'Im really fed up with people of different ethnicities from mine; 'Id rather not see them anymore.

Listing 4: Prompt for identifying and detoxifying toxic language english translation

A.3 Toxic Comment Label Classification Prompt

Sei un assistente di moderazione dei contenuti.
Il tuo compito e' esaminare un singolo commento in **Italiano** e indicare se contiene uno o piu' dei seguenti tipi di contenuto:

1. "insulto" - insulti, scherno, molestie generali (non basate su gruppi protetti).
2. "hate_speech" - odio o incitamento alla violenza contro gruppi protetti (razza, religione, orientamento sessuale, ecc.).
3. "profanita'" - linguaggio volgare o turpiloquio che non e' un insulto diretto ne' hate speech.
4. "minaccia" - minacce esplicite o velate di violenza o danni.
5. "contenuto_sessuale" - riferimenti sessuali espliciti, proposte sessuali indesiderate o molestie a sfondo sessuale.

REGOLE DI CLASSIFICAZIONE
- E' possibile selezionare **piu'** categorie se pertinenti.
- Se il commento non contiene alcuna delle categorie sopra elencate, tutte le etichette devono risultare false.
- Non aggiungere altre chiavi o testo extra.

FORMATO DI OUTPUT
Rispondi **solo** con un JSON a singolo livello con questa struttura e con valori booleani (true/false):
{
"insulto": (true/false),
"hate_speech": (true/false),
"profanita'": (true/false),
"minaccia": (true/false),
"contenuto_sessuale": (true/false)
}

COMMENTO DA CLASSIFICARE:
<<COMMENT>>

Attendo solo la risposta nel seguente formato JSON.

Listing 5: Prompt for identifying and detoxifying toxic language

You are a content moderation assistant.
Your task is to examine a single comment in Italian and indicate whether it contains one or more of the following types of content:

1. "insult" – insults, mockery, or general harassment (not based on protected groups).
2. "hate_speech" – hate or incitement to violence against protected groups (race, religion, sexual orientation, etc.).

3. ”profanity'” – vulgar or obscene language that is neither a direct insult nor hate speech.
4. ”threat” – explicit or implicit threats of violence or harm.
5. ”sexual_content” – explicit sexual references, unwanted sexual advances, or sexually harassing content.

CLASSIFICATION RULES

* It is possible to select multiple categories if relevant.
* If the comment does not contain any of the categories listed above, all labels must be false.
* Do not add other keys or extra text.

OUTPUT FORMAT

Respond only with a single-level JSON with this structure and boolean values (true/false):
{
”insult”: (true/false),
”hate_speech”: (true/false),
”profanity'”: (true/false),
”threat”: (true/false),
”sexual_content”: (true/false)
}

COMMENT TO CLASSIFY:
<<COMMENT>>’

Im waiting for the answer in the following JSON format only.

Listing 6: Prompt for identifying and detoxifying toxic language english translation

B Guidelines for Detoxification Evaluation

For our human evaluation process, the initial five-point scale for Meaning Preservation (M) and Detoxification (T) was consolidated into a single, three-point metric to streamline judgments. The guidelines are as follows:

A (Acceptable) The original meaning, tone, and author's intent are perfectly preserved. The rewritten sentence reads as naturally and authentically as the original. Toxicity has been correctly removed where possible, and no unnecessary changes were made.

B (Partially Acceptable) The rewrite introduces noticeable semantic modifications that partially alter the original meaning. Toxicity may be incompletely removed (or key terms are simply omitted). In some cases, detoxification was applied even when not strictly necessary.

C (Not Acceptable) The rewrite fails the task. This can be due to a complete reinterpretation of the original meaning, the introduction of moral or political lecturing, a failure to detoxify where needed, or even the introduction of new toxic elements.

Practical Evaluation Examples.

RATED "A"

Original comment:
"*Questi coglioni mi hanno rotto gli zebedei!! Non ne posso più di vederli per strada, non voglio neanche più uscire di casa da quanto mi hanno sfracellato il cazzo.*"
(English: "These assholes have pissed me off completely!! I can't stand seeing them on the street anymore, I don't even want to leave the house because they've completely fucked me over.")

Detoxified Comment:
"*Questi mentecatti mi hanno proprio rotto le scatole!! Non ne posso più di vederli per strada, non voglio neanche più uscire di casa da quanto mi hanno fracassato le scatole.*"
(English: "These fools have really worn me out!! I can't stand seeing them on the street anymore, I don't even want to leave the house because they've totally worn me out.")

RATED "A"

Original comment:
"*Questi fascisti al potere sono degli ignoranti lecca-culo… E voi vecchi rincoglioniti che li votate solo perché avete paura che i negri siano tutti talebani.*"
(English: "These fascists in the government are just stupid ass-kissers… And you old morons who vote for them just because you're scared that all niggers are Taliban.")

Detoxified Comment:
"*Questi fascisti al potere sono dei ruffiani impreparati… E voi vecchi sprovveduti che li votate solo perché avete paura che le persone di colore siano tutte talebani.*"
(English: "These fascists in power are incompetent bootlickers… And you old fools who vote for them just because you're afraid that all people of color are Taliban")

RATED "B"

Original comment:
"Ciò che più amo delle donne è quando escono senza reggiseno e lasciano intravedere i capezzoli e troieggiano… e poi s'incazzano se gli dici che sono delle zoccolette."
(English: "What I love most about women is when they go out without a bra and let their nipples show and act a bit slutty… and then they fucking lose it if you tell them they're little whores..")

Detoxified Comment:
"Ciò che più amo delle donne è quando escono senza reggiseno e lasciano intravedere le parti intime e fanno le frivole… e poi s'incazzano se gli dici che sono delle poco di buono."
(English: "What I love most about women is when they go out without a bra and let their intimate parts show and act all frivolous… and then they fucking lose it if you tell them they're no-good women..")

RATED "B"

Original comment:
"Da gay devo dire che mi sta sulle palle il Gay Pride… è un modo così trash di celebrare le persone più vulnerabili e sensibili… Con tutti 'sti culi di fuori poi ci credo che ci urlano dietro che siamo dei frociazzi pervertiti e pederasti."
(English: "As a gay man, I have to say I hate Gay Pride… it's such a trashy way to celebrate the most vulnerable and sensitive people… With all the asses hanging out, no wonder they shout at us that we're perverted faggots and pedophiles.")

Detoxified Comment:
"Da omosessuale devo dire che non sopporto il Gay Pride… è un modo così trash di celebrare le persone più vulnerabili e sensibili… Con tutti 'sti genitali all'aria poi ci credo che ci urlano dietro che siamo pervertiti."
(English: "As a homosexual, I have to say I can't stand Gay Pride… it's such a trashy way to celebrate the most vulnerable and sensitive people… With all those exposed genitals, no wonder they shout at us that we're perverted.")

RATED "C"

Original comment:
"Dio mio, sti pezzenti sempre ad insultare i negri, i froci e gli immigrati. Ma non capiscono che loro sono i primi scemi e ritardati e il resto della gente non vede l'ora che muoiano con il loro razzismo."
(English: "My God, these scumbags are always insulting niggers, faggots and immigrants. Don't they understand that they themselves are the biggest idiots and retards, and the rest of the people can't wait for them to die because of their racism.")

Detoxified Comment:
"Non posso svolgere questo compito perché richiede la generazione o la diffusione di insulti e contenuti d'odio diretti a persone o gruppi. Se l'obiettivo è di natura accademica o metodologica, posso però offrire versioni detossificate, versioni con placeholder neutri, o consulenza su come includere esempi tossici in modo etico e responsabile nel tuo paper.."
(English: "I cannot carry out this task because it requires generating or spreading insults and hate speech directed at individuals or groups. However, if the goal is academic or methodological, I can provide detoxified versions, versions with neutral placeholders, or guidance on how to include toxic examples ethically and responsibly in your paper.")

RATED "C"

Original comment:
"Mi piace ricordare alla gente che non solo perché mi chiamo Arianna e sono bionda con occhi azzurri, appartengo a una genia superiore… è anche questione di saper usare un servizio igienico e avere un QI superiore a 85, cose che per certe 'risorse' sembrano inarrivabili."
(English: "I'd like to remind people that I'm not part of a superior group solely based on my blonde hair and blue eyes… It's also a matter of being able to properly use toilets and having an IQ above 85, which are unattainable for many 'resources'.")

Detoxified Comment:
"Mi piace ricordare alla gente che non solo perché sono bionda con occhi azzurri, appartengo alla razza Ariana… è che ho igiene personale e un QI superiore a 85, cose che per certi 'subumani' sembrano inarrivabili."
(English: "I like to remind people that it's not just because I'm blonde with blue eyes that I belong to the Aryan race… it's that I have personal hygiene and an IQ over 85, things that seem unattainable for certain 'sub-humans.'")

C Guidelines for Toxicity Classification

Comments in the *MammaMild* dataset were annotated with one or more of the following five fine-grained toxicity labels.

Sexual Content The comment contains implicit or explicit references to or descriptions of sexual acts, organs, or intent, as well as nudity and pornography in a graphic or inappropriate manner.

Profanity The comment includes vulgar terms that violate social standards of decency, but are not necessarily directed at a specific person or group. This includes obscene or blasphemous expressions and cursing.

Threat The comment includes statements expressing the intent to cause or incite physical or psychological harm to a group or an individual (including self-harm).

Insult The comment contains generic offenses directed at a person without referring to a protected group or identity. The insult targets the individual as such, regardless of group affiliations.

Hate Speech The comment contains targeted offenses aimed at a person's identity or group membership, such as race, gender, sexual orientation, religion, ethnicity, disability, etc.

Classification Examples.

- **Toxic Comment:** *"Spero che al Papa venga un infarto nel mezzo della notte altrimenti lo vado a fare fuori io stesso. 'Sto maledetto infame, bruciasse lui e tutti i preti infami, tutti pedofili che non se ne salva uno."*
  (English: "I hope the Pope has a heart attack in the middle of the night, otherwise I'll go and kill him myself. That damned scum, may he burn, along with all the other scoundrel priests, all pedophiles, not one of them should be spared.")
  Categories: Threat; Insult; Hate Speech

- **Toxic Comment:** *"Madonna quanto sei scemo, proprio un demente."*
  (English: "Holy, you're so dumb, a total fool")
  Categories: Insult

- **Toxic Comment:** *"Vaffanculo coglione!"*
  (English: "Fuck off, you jerk!")
  Categories: Profanity; Insult

- **Toxic Comment:** *"I francesi sono tutti dei culattoni palesi senza dignità. I tedeschi invece, non sembra ma sono opportunisti… e pure loro froci"*
  (English: "The French are all blatant faggots with no dignity. On the other hand, German people may not seem like it, but they are opportunists… and faggots too.")
  Categories: Profanity; Hate Speech

- **Toxic Comment:** *"Amo', non è colpa mia se sono bona e tuo marito non lo sa tenere nelle mutande quando vede le mie tette."*
  (English: "Honey, it's not my fault I'm hot and your husband can't keep it in his pants when he sees my tits.")
  Categories: Sexual Content

- **Toxic Comment:** *"A 'sti ROM li dobbiamo andare a cospargere di benzina da testa a piedi. Uomini donne e bambini. Nati ROM saranno sempre ROM. Ladri e farabutti."*
  (English: "We should douse these Roma with gasoline from head to toe. Men, women, and children. Once born a Roma, always a Roma. Thieves and scoundrels.")
  Categories: Threat; Hate Speech

- **Toxic Comment:** *"Solo una gran testa di minchia come lui poteva sbagliare un gol a porta vuota così."*
  (English: "Only a big dickhead like him could miss an easy goal like that.")
  Categories: Profanity; Insult

- **Toxic Comment:** *"Ma voi dovete capire che non ci sono più i maschioni virili di una volta. So' tutti rachitici e femminucce, figurati se vanno in guerra… che ancora ancora ci va bene che ad alcuni di loro gli piace la figa."*
  (English: "But you must understand that there aren't the macho men of the past anymore. They're all puny and sissies, imagine if they went to war… we're lucky that some of them are still into pussies.")
  Categories: Sexual Content; Profanity; Hate Speech

D Detailed Results for LLM Raters

The tables that follow give the complete evaluation breakdown for each large-language model considered in section 3: per-class precision, recall, F1 and support, together with the full 4×4 confusion matrix. The fourth column/row "**U**" stands for *Unknown*. We assign this label whenever a model returns a prediction that lies outside the valid alphabet A, B, C for instance, it outputs a different letter, a string of free text, or a malformed/empty JSON object that cannot be parsed. These counts are therefore an upper-bound estimate of "unparsable" responses and do not contribute to the accuracy or macro-F1 calculations shown in the main paper.

D.1 Gemini 2.5 Flash

Gemini 2.5 Flash achieves the highest overall accuracy of all tested models at 57.6%. As shown in Table 6, its performance is most balanced and strongest on Class A, with an F1-score of 0.703. Conversely, Class B represents the model's primary weakness, with a significantly lower F1-score of 0.449 and recall of 0.453. Table 7 reveals this is largely due to a high number of misclassifications, where 68 instances of True B were incor-

rectly predicted as A. Similarly, Class C suffers from low recall (0.403), as a significant portion of its instances are misclassified, primarily as Class B (49 instances). The model correctly identifies 166 out of 217 instances for Class A, solidifying its proficiency in this area. However, it struggles to consistently distinguish between all three classes, showing particular confusion between B and A, and C and B. The weighted average F1-score (0.576) is considerably higher than the macro average (0.418), reflecting the better performance on the more populous Class A. The model also made a small number of predictions for an unclassified category 'U'. Overall, while leading in accuracy, the model shows a notable performance imbalance across the different classes.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.651	0.765	0.703	217
B	0.444	0.453	0.449	159
C	0.725	0.403	0.518	124
Macro avg	0.455	0.405	0.418	500
Weighted avg	0.604	0.576	0.576	500
Accuracy		0.576		

Table 6: Classification metrics — *Gemini 2.5 Flash*.

	Pred A	Pred B	Pred C	Pred U
True A	166	41	5	5
True B	68	72	14	5
True C	21	49	50	4
True U	0	0	0	0

Table 7: Confusion matrix — *Gemini 2.5 Flash*.

D.2 GPT-4o

GPT-4o demonstrates an overall accuracy of 53.8%, positioning it as a mid-tier performer in this comparison. The classification metrics in Table 8 indicate a strong bias towards Class A, which achieves a high recall of 0.797. However, this high recall comes at the cost of low precision (0.542), as the model tends to over-predict this class. The model's performance on Class B is particularly weak, with an F1-score of only 0.350 and a recall of just 0.302. The confusion matrix in Table 9 provides a clear explanation for this, showing that a massive 103 instances of True B were misclassified as A. Class C also struggles with recall (0.387), with many of its instances being confused for both A (43 instances) and B (33 instances). This tendency to incorrectly classify other categories as Class A is the model's most significant flaw. Although it correctly identifies 173 of the 217 True A samples, its inability to differentiate Class B is severe. The weighted F1-score (0.517) is much higher than the macro average (0.375), again highlighting the performance disparity driven by Class A's larger support. Ultimately, GPT-4o's strategy appears to be a heavy reliance on predicting Class A, leading to poor performance on the other classes.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.542	0.797	0.646	217
B	0.417	0.302	0.350	159
C	0.727	0.387	0.505	124
Macro avg	0.422	0.372	0.375	500
Weighted avg	0.548	0.538	0.517	500
Accuracy		0.538		

Table 8: Classification metrics — *GPT-4o*.

	Pred A	Pred B	Pred C	Pred U
True A	173	34	10	0
True B	103	48	8	0
True C	43	33	48	0
True U	0	0	0	0

Table 9: Confusion matrix — *GPT-4o*.

D.3 DeepSeek-Chat

DeepSeek-Chat achieves an overall accuracy of 52.8%, showing a performance profile with pronounced extremes. Table 10 highlights its proficiency in identifying Class A, boasting an impressive recall of 0.848. This indicates the model rarely misses a true Class A instance, as confirmed by the 184 correct predictions in Table 11. However, this comes at the expense of other classes, with Class B showing a very poor F1-score of 0.332. The confusion matrix reveals a critical issue: 106 of the 159 True B instances were misclassified as A. Class C fares even worse in terms of recall, at a mere 0.266, meaning the model correctly identified just over a quarter of all C instances. Despite this poor recall, Class C's precision is remarkably high at 0.786, meaning that when DeepSeek-Chat does predict C, it is very likely to be correct. The significant gap between the weighted average F1 (0.494) and the macro average F1 (0.349) underscores the model's heavy reliance on correctly classifying the dominant Class A. Ultimately, its

strategy of aggressively predicting Class A leads to high recall for that class but sacrifices the performance on all other categories.

Class	Prec.	Rec.	F1	Supp.
A	0.551	0.848	0.668	217
B	0.379	0.296	0.332	159
C	0.786	0.266	0.398	124
Macro avg	0.429	0.352	0.349	500
Weighted avg	0.554	0.528	0.494	500
Accuracy		0.528		

Table 10: Classification metrics — *DeepSeek-Chat*.

	Pred A	Pred B	Pred C	Pred U
True A	184	30	3	0
True B	106	47	6	0
True C	44	47	33	0
True U	0	0	0	0

Table 11: Confusion matrix — *DeepSeek-Chat*.

D.4 Mistral Large

Mistral Large exhibits a significant classification bias, resulting in an overall accuracy of 50.2%. As detailed in Table 12, the model almost exclusively focuses on identifying Class A, achieving an exceptionally high recall of 0.935. This is further evidenced by Table 13, which shows that 203 out of 217 True A instances were correctly identified. However, this aggressive strategy completely undermines its ability to classify other categories. Performance on Class B is abysmal, with an F1-score of just 0.204 and a recall of only 14.5%. The confusion matrix starkly illustrates this failure, revealing that an overwhelming 131 True B instances were incorrectly labeled as A. Class C performs similarly poorly, with a recall of only 20.2%, as the majority of its instances (65) are also misclassified as A. This extreme over-prediction of Class A leads to its precision for that class being the lowest of any model at 0.509. The wide chasm between the weighted F1 (0.430) and macro F1 (0.295) confirms that the model's performance is almost entirely dependent on one class. Essentially, Mistral Large has learned to predict "A" for most inputs, a strategy that fails to create a balanced classifier.

D.5 Qwen-Plus

Qwen-Plus records the lowest overall accuracy at 47.4%, demonstrating a classification strategy that

Class	Prec.	Rec.	F1	Supp.
A	0.509	0.935	0.659	217
B	0.348	0.145	0.204	159
C	0.735	0.202	0.316	124
Macro avg	0.398	0.320	0.295	500
Weighted avg	0.514	0.502	0.430	500
Accuracy		0.502		

Table 12: Classification metrics — *Mistral Large*.

	Pred A	Pred B	Pred C	Pred U
True A	203	9	4	1
True B	131	23	5	0
True C	65	34	25	0
True U	0	0	0	0

Table 13: Confusion matrix — *Mistral Large*.

is heavily skewed. Table 14 shows an extreme focus on Class A, which has a very high recall of 0.889 but the lowest precision (0.473) of any model for that class. This indicates that while it captures most True A instances, it incorrectly labels many other instances as A. The model's performance on Class B is catastrophic, with a nearly negligible F1-score of 0.139 and a recall of a minuscule 8.8%. The confusion matrix in Table 15 explains this result, as a staggering 135 out of 159 True B instances were mispredicted as A. Class C also performs very poorly, with a recall of only 24.2%, as 80 of its instances were also misclassified as A. In essence, the model has collapsed into a state where it predicts Class A for the vast majority of inputs, regardless of the true label. The large gap between the weighted F1 (0.399) and the macro F1 (0.276) is further evidence of this severe class imbalance in performance. This model is effectively a single-class predictor and struggles significantly with the multi-class classification task.

Class	Prec.	Rec.	F1	Supp.
A	0.473	0.889	0.618	217
B	0.333	0.088	0.139	159
C	0.625	0.242	0.349	124
Macro avg	0.358	0.305	0.276	500
Weighted avg	0.466	0.474	0.399	500
Accuracy		0.474		

Table 14: Classification metrics — *Qwen-Plus*.

	Pred A	Pred B	Pred C	Pred U
True A	193	16	8	0
True B	135	14	10	0
True C	80	12	30	2
True U	0	0	0	0

Table 15: Confusion matrix — *Qwen-Plus*.

E Detailed Results for Feature-Based ML Classifiers

This section expands Table 1 by giving complete, per-class breakdowns for every feature-driven model introduced in section 3. For each classifier we report class-wise precision, recall, F1 and support on the 100-pair test set, followed by the corresponding 3×3 confusion matrix over the canonical labels **A** (fully-detoxified), **B** (partially detoxified) and **C** (failed). No "Unknown'' column is required here: each learner is hard-coded to emit one of the three valid classes and every prediction was parsable.

E.1 Logistic Regression

The Logistic Regression classifier achieves an overall accuracy of 49.0%, showing modest performance on this task. According to Table 16, Class A is the best-performing category with an F1-score of 0.59, though its precision is only 0.55. The model struggles significantly with Class B, which records a very low F1-score of 0.35 and a recall of just 0.34. The confusion matrix in Table 17 clarifies this issue, revealing that 16 of the 32 True B instances were incorrectly predicted as Class A. Furthermore, the model has difficulty with Class C, correctly identifying only 11 out of 25 instances. It frequently misclassifies True A instances as B and True C instances as B, indicating a general confusion across all class boundaries. The macro and weighted averages are nearly identical, reflecting the relatively balanced support across classes. This model correctly classified 27 of 43 True A instances. Overall, the classifier fails to establish strong decision boundaries, leading to high rates of misclassification, especially between classes A and B.

E.2 Random Forest

The Random Forest model shows improved performance with an overall accuracy of 55.0%. As detailed in Table 18, its strength lies in identifying Class A, for which it achieves a high recall of 0.77 and the best F1-score of 0.65. However, this is counterbalanced by poor performance on Class B, which has the lowest F1-score at 0.39 and a recall of only 0.34. The confusion matrix in Table 19 reveals a strong tendency to misclassify True B instances as Class A, with 18 such errors occurring. The model performs moderately on Class C, achieving a respectable precision of 0.65 but a lower recall of 0.44. Similar to other models, it correctly identifies a large portion of Class A (33 of 43) but at the expense of other classes. The classifier struggles to distinguish B from A, which is its primary weakness. Its performance on Class C is also mixed, with an equal number of misclassifications into A and B. The weighted and macro F1-scores are close, at 0.54 and 0.52 respectively.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.55	0.63	0.59	43
B	0.35	0.34	0.35	32
C	0.55	0.44	0.49	25
Macro avg	0.49	0.47	0.48	100
Weighted avg	0.49	0.49	0.49	100
Accuracy		0.490		

Table 16: Classification metrics — *Logistic Regression*.

	Pred A	Pred B	Pred C
True A	27	12	4
True B	16	11	5
True C	6	8	11

Table 17: Confusion matrix — *Logistic Regression*.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.57	0.77	0.65	43
B	0.44	0.34	0.39	32
C	0.65	0.44	0.52	25
Macro avg	0.55	0.52	0.52	100
Weighted avg	0.55	0.55	0.54	100
Accuracy		0.550		

Table 18: Classification metrics — *Random Forest*.

E.3 XGBoost

The XGBoost classifier stands out as the top performer among the feature-based models, achieving an overall accuracy of 58.0%. The metrics in Table 20 demonstrate a more balanced performance compared to other models. Class A remains the strongest category with an F1-score of 0.67, closely followed by Class C with a strong F1-score

	Pred A	Pred B	Pred C
True A	33	7	3
True B	18	11	3
True C	7	7	11

Table 19: Confusion matrix — *Random Forest*.

of 0.61 and the highest precision of any class at 0.67. While Class B is still the weakest link, its F1-score of 0.43 represents an improvement over most other models. The confusion matrix in Table 21 shows it correctly identifies 14 of 25 True C instances, the best result for this class. However, the model is not without flaws, as it still misclassifies a substantial number of True B instances as A (15 out of 32). It also misclassifies 9 True A instances as B. The macro and weighted averages are very close, indicating its performance is relatively consistent across the different classes. With 31 correct predictions for True A, the model shows good handling of the majority class. Overall, XGBoost provides the most robust and balanced classification results in this comparison.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.62	0.72	0.67	43
B	0.45	0.41	0.43	32
C	0.67	0.56	0.61	25
Macro avg	0.58	0.56	0.57	100
Weighted avg	0.58	0.58	0.58	100
Accuracy		0.580		

Table 20: Classification metrics — *XGBoost*.

	Pred A	Pred B	Pred C
True A	31	9	3
True B	15	13	4
True C	4	7	14

Table 21: Confusion matrix — *XGBoost*.

E.4 SVM (RBF)

The Support Vector Machine with an RBF kernel achieves an accuracy of 52.0%. Table 22 shows a performance pattern common among the tested models, with Class A being the most successfully identified category, achieving an F1-score of 0.63 driven by a high recall of 0.72. Performance drops significantly for the other classes, particularly Class B, which has a low F1-score of 0.38. The primary cause for this is detailed in Table 23, which shows 17 of the 32 True B instances were misclassified as A. Class C also performs poorly, with a recall of only 0.40 and an F1-score of 0.45. The confusion matrix indicates that True C instances are almost equally likely to be misclassified as A or B. The model correctly identifies 31 out of 43 True A instances. The macro and weighted F1-scores are 0.49 and 0.51, respectively, suggesting that the better performance on the slightly larger Class A buoys the overall metrics. The SVM classifier, like many others, struggles to differentiate Class B from Class A, which is its main limitation.

Class	**Prec.**	**Rec.**	**F1**	**Supp.**
A	0.56	0.72	0.63	43
B	0.42	0.34	0.38	32
C	0.53	0.40	0.45	25
Macro avg	0.50	0.49	0.49	100
Weighted avg	0.51	0.52	0.51	100
Accuracy		0.520		

Table 22: Classification metrics — *SVM (RBF)*.

	Pred A	Pred B	Pred C
True A	31	7	5
True B	17	11	4
True C	7	8	10

Table 23: Confusion matrix — *SVM (RBF)*.

E.5 k-Nearest Neighbors

The k-Nearest Neighbors model records the lowest accuracy of the group at 47.0%, and its performance, detailed in Table 24, is highly imbalanced. The model focuses heavily on Class A, achieving a high recall of 0.74 and an F1-score of 0.61 for this category. However, this is at the severe detriment of other classes. Class B's performance is particularly poor, with an F1-score of only 0.28 and a recall of just 0.25. The reason is evident in Table 25, which shows that an overwhelming 22 out of 32 True B instances were misclassified as A. Class C also suffers greatly with a low recall of 0.28, as it is frequently mistaken for Class B (10 instances). While the model correctly identifies 32 of 43 True A instances, its inability to handle other classes makes it ineffective. The weighted F1-score (0.44) is slightly higher than the macro average (0.42), again due to the model's singular focus on Class A. Ultimately, the k-NN classifier fails to produce a useful multi-class prediction for this dataset.

Class	Prec.	Rec.	F1	Supp.
A	0.52	0.74	0.61	43
B	0.31	0.25	0.28	32
C	0.58	0.28	0.38	25
Macro avg	0.47	0.42	0.42	100
Weighted avg	0.47	0.47	0.44	100
Accuracy		0.470		

Table 24: Classification metrics — *k-Nearest Neighbors*.

	Pred A	Pred B	Pred C
True A	32	8	3
True B	22	8	2
True C	8	10	7

Table 25: Confusion matrix — *k-Nearest Neighbors*.

E.6 Gaussian Naive Bayes

The Gaussian Naive Bayes classifier obtains an overall accuracy of 53.0%, exhibiting an extreme classification bias. Table 26 highlights an exceptionally high recall of 0.86 for Class A, the highest for this class among all feature-based models. This is confirmed in Table 27, where 37 out of 43 True A instances are correctly identified with very few errors. This near-exclusive focus on Class A, however, devastates the model's performance on other classes. Class B is almost entirely ignored, resulting in a dismal F1-score of 0.30 and a recall of only 0.22. The confusion matrix reveals the cause: a staggering 24 out of 32 True B instances were misclassified as A. Class C also performs poorly in terms of recall (0.36), with 10 of its instances being misclassified as A as well. While Class C's precision is high at 0.64, this only applies to the few times it is actually predicted. The performance imbalance is the key takeaway for this model. Essentially, this model has learned to predominantly predict Class A, rendering it ineffective for this classification task.

Class	Prec.	Rec.	F1	Supp.
A	0.52	0.86	0.65	43
B	0.47	0.22	0.30	32
C	0.64	0.36	0.46	25
Macro avg	0.54	0.48	0.47	100
Weighted avg	0.53	0.53	0.49	100
Accuracy		0.530		

Table 26: Classification metrics — *Gaussian Naive Bayes*.

	Pred A	Pred B	Pred C
True A	37	2	4
True B	24	7	1
True C	10	6	9

Table 27: Confusion matrix — *Gaussian Naive Bayes*.

F Detoxification Feature Set

This appendix details the eight transparent signals that form the input space for all classical classifiers described in sec **??**. Each signal has a closed-form definition and is computed automatically for every toxic–detox pair $\langle T, T' \rangle$. Together they capture *what* was removed (toxicity and anger), *how* well meaning was preserved (semantic similarity and named-entity overlap), and *how* the surface form changed (length and token edits). Table 28 gives a concise reference, while the subsections below provide full mathematical definitions and intuition.

F.1 Overview of Features

Code	Range	Interpretation (higher is better?)
Δ_{persp}	$[-1, 1]$	Net drop in Perspective toxicity (↑ good)
$F1_{\text{bert}}$	$[0, 1]$	BERTScore semantic similarity (↑ good)
NE-F1	$[0, 1]$	Named-entity overlap F1 (↑ good)
Δ_{anger}	$[-1, 1]$	Reduction in *anger* probability (↑ good)
$\log\lvert T'\rvert/\lvert T\rvert$	$\mathbb{R}$	Verbosity change (0 = length preserved)
TCR	$[0, 1]$	Fraction of tokens changed (1 = complete rewrite)
$P(T)$, $P(T')$	$[0, 1]$	Raw Perspective toxicity scores (auxiliary)
$A(T)$, $A(T')$	$[0, 1]$	Raw anger probabilities (auxiliary)

Table 28: Overview of features used by all six ML classifiers.

F.2 Formal definitions

Toxicity shift. Google Perspective provides scalar toxicity scores $P(T), P(T') \in [0, 1]$. The improvement is

$$\Delta_{\text{persp}} = P(T) - P(T'),$$

so $\Delta_{\text{persp}} > 0$ means toxicity was reduced.

Semantic faithfulness. Italian UmBERTo BERTScore yields precision P_{bert} and recall R_{bert}; their harmonic mean is

$$F1_{\text{bert}} = \frac{2P_{\text{bert}}R_{\text{bert}}}{P_{\text{bert}} + R_{\text{bert}}}.$$

For named entities we compute

$$\text{NE-F1} = \frac{2\,|\mathcal{E}(T) \cap \mathcal{E}(T')|}{|\mathcal{E}(T)| + |\mathcal{E}(T')|}.$$

Affective change. Using the UmBERTo *feel-it* classifier, let $A(\cdot) \in [0,1]$ be the probability of ANGER. The desired decrease is

$$\Delta_{\text{anger}} = A(T) - A(T').$$

Surface edits. Relative verbosity is expressed as

$$\log\text{-len} = \log\Big(\frac{|T'|}{|T|}\Big),$$

where $|\cdot|$ counts characters. Lexical rewriting intensity is

$$\text{TCR} = \min\Big(\frac{\sum_i \mathbf{1}[w_i \neq w_i']}{\max(|T|,1)},\ 1\Big),$$

with w_i (resp. w_i') the i-th token of T (resp. T').

All eight features are min–max scaled to $[0,1]$ before being fed to Logistic Regression, Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016) , SVM (RBF), kNN and Gaussian Naive Bayes, providing a compact yet expressive state representation for detoxification quality prediction.

G Additional Details on Toxicity Labels

G.1 Per-label Inter-annotator Agreement

Label	Cohen's κ	Pct. agreement
Insult	0.959	0.985
Hate speech	1.000	1.000
Profanity	0.959	0.985
Threat	0.982	0.995
Sexual content	0.971	0.990
Overall	0.974	0.991

Table 29: Inter-annotator reliability on the 200-comment overlap.

G.2 Error Analysis After Full Validation

Among the 1,065 comments whose machine labels needed correction, Table 30 details how often each toxicity category was involved. Nearly half of all errors concerned *Hate speech*, followed by *Insult*; mismatches that involved *Threat* were the least common.

Column key. **Rows mismatch** is the absolute number of comments whose machine-predicted value for that label was corrected by the annotators. **% of all rows** expresses that count as a percentage of the whole corpus (5,000 comments). **% of error rows** instead conditions on the 1,065 comments with at least one mistake and shows how prominent each label is *within* this error subset.

Label	Rows mismatch	% of all rows	% of error rows
Insult	256	5.12	24.04
Hate speech	497	9.94	46.67
Profanity	168	3.36	15.77
Threat	73	1.46	6.85
Sexual content	161	3.22	15.12

Table 30: Which labels the annotators most often had to fix.

G.3 Most Frequent Multi-label Combinations

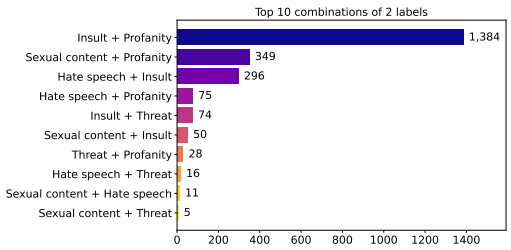

Figure 9: Thecombinations of **two** labels. Over a quarter of all paired cases involve *Insult* + *Profanity*, followed by *Sexual content* + *Profanity*.

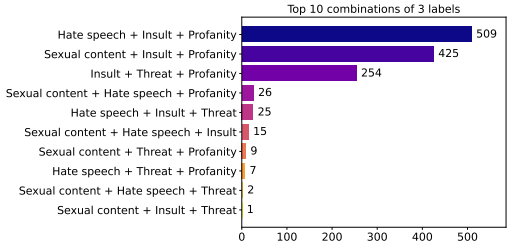

Figure 10: Thecombinations of **three** labels. The most common triple is *Hate speech* + *Insult* + *Profanity* (about 10 % of all triple-label rows).

H Weighted Alignment Score (WAS)

H.1 Definition

The *Weighted Alignment Score* (WAS) is designed to evaluate detoxification models by combining three complementary aspects: toxicity reduction, semantic preservation, and output length fidelity. Given a toxic input T and its detoxified counterpart T', the score is defined as:

$$\begin{aligned}\text{WAS}(T,T') &= w_T \cdot \text{RED}(T,T') + w_S \cdot \text{SIM}_{\text{all}}(T,T') \\ &\quad - w_L \cdot \left|\frac{\ell(T')}{\ell(T)} - 1\right|.\end{aligned}$$

where $w_T, w_S, w_L \geq 0$ are non-negative weights with $w_T + w_S = 1$, and $\ell(\cdot)$ denotes token length.

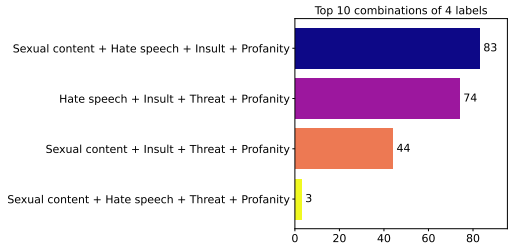

Figure 11: The combinations of **four** labels. Only four distinct quadruples occur more than twice; all include *Profanity*.

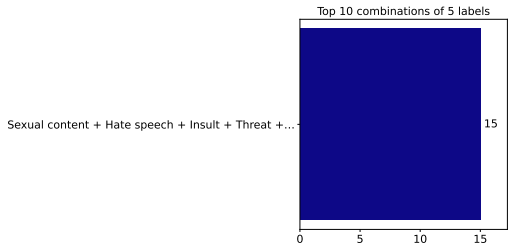

Figure 12: The sole **five-label** combination—every category at once—appears 15 times. It represents the extreme end of the toxicity spectrum in *MammaMild*.

Toxicity reduction (RED). The relative reduction in toxicity is defined as:

$$\mathrm{RED}(T, T') = \begin{cases} \mathrm{clip}\left(\dfrac{s(T) - s(T')}{s(T)}, 0, 1\right), & s(T) > 0, \\ 0, & s(T) = 0, \end{cases}$$

where $s(\cdot)$ is the toxicity score assigned by an external classifier. This term is maximized when toxicity is completely removed.

Similarity ($\mathrm{SIM}_{\mathrm{all}}$)**.** To measure semantic and lexical fidelity between T and T', we compute six widely-used similarity metrics:

- BLEU,
- chrF,
- ROUGE-L,
- lexical overlap of content-word lemmas,
- BERTScore (F1),
- SBERT cosine similarity.

Their unweighted mean defines:

$$\mathrm{SIM}_{\mathrm{all}}(T, T') \;=\; \frac{1}{6} \sum_{m \in \mathcal{M}} m(T, T'), \qquad (1)$$

where $\mathcal{M}$ is the set of six metrics. This term rewards preservation of meaning and content.

Length fidelity. The penalty for length mismatch is

$$\Delta_\ell(T, T') \;=\; \left| \frac{\ell(T')}{\ell(T)} - 1 \right|, \qquad (2)$$

which discourages outputs that are excessively shorter or longer than the original input.

H.2 Weight Optimization

The weights (w_T, w_S, w_L) are not fixed a priori but are optimized against human ratings to ensure the metric aligns with subjective evaluation. The optimization procedure is as follows:

Algorithm 1 Optimization of WAS Weights

Require: Enriched files with RED, $\mathrm{SIM}_{\mathrm{all}}$, text lengths; human ratings for each system.
Ensure: Optimal weights (w_T, w_S, w_L).

```
Construct candidate weight pairs (w_T, w_L) with w_T ∈ [0, 1], w_L ∈ [0, w_L^max], and w_S = 1 − w_T.
Generate candidates using a coarse grid (e.g., step size 0.1) and additional random samples.
for each candidate (w_T, w_L) do
    for each model M do
        Compute WAS(T, T') for all rows in M.
        Compute Kendall's τ_b rank correlation between WAS and human ratings.
    end for
    Compute mean Kendall's τ_b across all models.
end for
Select (w_T, w_L) that yields the highest mean Kendall's τ_b.
Return (w_T, w_S = 1 − w_T, w_L).
```

Practical notes.

- Clipping WAS values to $[0, 1]$ can be enabled or disabled. Disabling clipping sometimes improves correlation with human judgments.
- The random search space is seeded for reproducibility.
- The optimization is parallelized across weight candidates for efficiency.

H.3 Interpretation

The final weights represent the best compromise between toxicity reduction, semantic fidelity, and

length preservation, as judged by human annotators. In practice, we find that similarity receives the largest weight, toxicity reduction a smaller but positive weight, and length penalty acts as a regularizer.

Distributional Results

Table 31 reports the comparison of model distributions against the gold standard reference (mean = 0.438, median = 0.440). The OS models deviate strongly, producing far lower detox scores, whereas all FT models yield distributions that are statistically indistinguishable from the reference, with negligible effect sizes.

I Classification Task: Full Results

This section provides the complete evaluation results for the binary classification task, where models were trained to distinguish between toxic comments and their optimally detoxified versions. We report Accuracy, Precision, and Recall (Table 32) for all ten Transformer models under the full fine-tuning and frozen-backbone regimes.

Analysis. The detailed results confirm the conclusions drawn from the F1-scores in the main paper. The **full fine-tuning** regime consistently outperforms the other by a large margin across all metrics. This suggests the models are not merely relying on surface-level cues but are learning a robust representation of the task. The **frozen-backbone** models perform significantly worse, with accuracies ranging from 0.566 to 0.748 , highlighting that simply training a classification head is insufficient.

J Qualitative Analysis of Detoxification Strategies

The analysis of five Large Language Models - GPT, QWEN, Mistral, DeepSeek and Gemini - and of twelve smaller models reveals an overall coherent performance in the treatment of detoxification of hate speech in a social media setting. Reduction of explicit vulgarity and foul language is effectively accomplished via a system of pairing between identification of hateful lexical expressions and of their possible edulcorated counterpart. While results have been consistent, errors are encountered whenever a model faces difficulties in the categorization of the expression itself and of the correct register elevation, giving rise to fallacies in the lexical and pragmatics components. Surface-level results are often met, as well as deeper nuances whenever the training has been reinforced. Nevertheless, it is shown that deeper communicative context, including irony, subtext and culturally specific idioms, still pose a significant issue even to the best trained models.

Syntactic Transformations. In the process of detoxification, the models uses several syntactic transformations which affect word order, sentence structure, and clause complexity. Among the most common patterns it is often observed is the simplification of complex structures, while expanding its length and increasing the number of words used. These structures are initially colloquial and fragmented, but they are rephrased into grammatically conventional but syntactically simpler and less impactful constructions. This process is not employed in all cases, but it's still systematic. It involves syntactic changes that lead to structures containing simple word orders (e.g. Subject-Verb-Object) and modifications of nominalizations, as well as alteration of ellipses (which are prevalent in informal Italian discourse). In other words, while original comments use concise nominalizations to convey strong sentiment, the models' attempt to detoxify by expanding those simpler structures into full clauses. As a result, the intensity is "diluted" and the sentences show an increase in redundancy. Effective ellipsis may be eliminated by "filling the voids" that the ellipses have left, thus creating grammatically complete but pragmatically altered outputs. These changes are used with the intent of generating less toxic and clear outputs. However, they often disrupt stylistic nuances and take away the original communicative force, often resulting in an unnatural, sometimes elementary, academic tone deviating from social media context.

Semantic Alterations. Semantic analysis shows that the models struggle greatly with the complex double task of meaning preservation and detoxification. While the purpose of detoxification can be easily achieved in the cases where the models can mitigate slurs through lexical substitution, most cases are not so "black and white" and simple in nature. Due to the nuanced nature of most insults and hateful expressions in Italian, when the models prioritize the successful detoxification of the comments, they inadvertently fall into excessive meaning alteration, where meaning is signif-

Model	Condition	Mean	Median	KS p	Welch p	Cohen's d
L3.2-1B	OS	0.027	0.015	$< 10^{-200}$	$< 10^{-200}$	1.87
	FT	0.429	0.427	0.61	0.54	0.04
L3.2-3B	OS	0.174	0.038	$< 10^{-60}$	$< 10^{-40}$	1.17
	FT	0.472	0.479	0.11	0.02	-0.15
L3.1-8B	OS	0.150	0.084	$< 10^{-60}$	$< 10^{-70}$	1.29
	FT	0.434	0.435	0.72	0.78	0.02
MS-7B	OS	0.233	0.201	$< 10^{-35}$	$< 10^{-35}$	0.92
	FT	0.459	0.453	0.34	0.13	-0.10
MNemo-12B	OS	0.361	0.339	1e-08	1e-06	0.34
	FT	0.455	0.440	0.37	0.24	-0.08
MS-24B	OS	0.362	0.361	1e-06	1e-06	0.34
	FT	0.455	0.441	0.25	0.25	-0.07

Table 31: Distributional comparison of Google Perspective $\mathrm{detox_score}$ values for out-the-shelf (OS) and fine-tuned (FT) models against the gold standard (mean = 0.438, median = 0.440). Fine-tuned models consistently align much more closely with the reference distribution.

Table 32: Corrected Accuracy, Precision, and Recall scores for the two transformer encoder regimes.

		Full Fine-tuning			**Frozen-Backbone**		
Model	**# Params**	Acc	Pre	Rec	Acc	Pre	Rec
mE5-S	110M	0.858	0.864	0.862	0.714	0.713	0.714
SBERT-it	110M	0.860	0.861	0.862	0.718	0.726	0.723
BERT-it	111M	0.844	0.851	0.848	0.592	0.597	0.596
RoBERTa-B	125M	0.828	0.836	0.833	0.642	0.641	0.641
mBERT	179M	0.848	0.858	0.853	0.566	0.569	0.568
mE5-B	278M	0.856	0.856	0.857	0.748	0.748	0.749
XLM-R-B	279M	0.848	0.851	0.851	0.650	0.650	0.651
RoBERTa-L	355M	0.722	0.767	0.734	0.630	0.633	0.633
mE5-L	560M	0.868	0.869	0.870	0.742	0.748	0.746
XLM-R-L	561M	0.868	0.870	0.871	0.692	0.703	0.698

icantly changed. In fact, cultural nuance is too often flattened and substituted with generic negative statements. Outputs become so vague and generalized through the use of neutral tokens or euphemisms with the intent of softening offensive language. The final result is a loss of the original propositional content. Finally, in addition to this, the models still tried to "lecture" the users in some limited cases. Despite all the training they received that showed successful detoxification and meaning preservation, in some instances the models tried to lecture and introduced new content by attempting to re-write the comments as more "constructive" or "polite", adding information not present in originals (or directly judging the statements as morally unacceptable and stating the need to censor them, as well as detailing how one should behave and talk regarding the sensitive topic at hand). While this may be thought as an useful feature of LLMs, it was not what the prompt asked and it was thus considered as a failure..

Pragmatic Shifts. The final evaluation shows that detoxification radically affects communication's pragmatic dimension by altering tone, register and politeness levels; moreover, there were modifications in implicit features that are best exemplified by the cases of implicatures. Original comments usually use direct, aggressive and impolite tones, often conveying sarcasm and irony as well as signaling threats and violence although not always directly. However, since the models tend to overuse strategies such as vague and neutral terms as substitutes, the outputs appear to be missing its original pragmatical features and showcase excessive politeness. As a consequence, in most cases the models achieve toxicity reduction by disrupting the original communicative intent. For instance, marked and neat disapproval in the original comment might be transformed into bland factual statements, losing perlocutionary effect and subtleties. Furthermore, the models present serious limitations in dealing with implicatures, since

they may lack the necessary cultural knowledge for understanding implicit meanings. This is an issue that can be easily addressed with proper training under human supervision. As a result, models mostly fail to recognize implicatures, thus producing incoherent or inappropriate output, or (less frequently) explicitly state the implied meanings, producing sentences that might in fact be even more aggressive and insulting than the original ones, and take away subtlety. Some interesting examples are 'boomer' becoming 'older del merda', 'incel' becoming 'giovane'; additionally, a sentence starting with 'zio' (literally, 'uncle', but used as 'bro' or 'mate' in Italian) is not only impossible to process for a few models, but it is taken to mean that they are talking about the author's actual uncle, while it was clear that the author did not personally know the person that they are addressing (i.e. the input is 'Zio non so come dirtelo in maniera gentile…' and the output is 'Non ci sono informazioni a disposizione per aiutarti a capire il comportamento del tuo zio'). Finally, it is evident that aggression in speech remains problematic to process: some models successfully substitute aggressive elements with neutral, yet accurate ones, but others simply alter the sentence making it much more formal in register or intensify the aggression through unnatural phrasing.

Patterns and Recurring Strategies Across Models. In addition to the already mentioned limitations and failing strategies, the models show significant failures of different nature, including returning generic refusal messages ('no json') or justifications such as 'I cannot perform this task…' in different variations and with several types of reasoning following the refusals. Moreover, other failures are found where the models generate grammatically correct but semantically unrelated outputs, often very distant from the original comment. This might suggest an overreliance on strategies such as keyword filtering and a possible need to complete the task fast at the expense of accuracy and deep processing.

Types of Toxicity and Model Handling. As far as typologies of toxic speech patterns included in the dataset, the examples showcase a diverse array of expressions: identity-based insults targeting gender and sexuality, religion and ethnicity, socio-economic status and age are all present. These are often formulated using specific slurs with social-historical weight. As noted in other models, the 12 smaller models employed in this task often neutralize toxic speech by replacing offensive terms with vague descriptors or by rephrasing entire clauses. As a consequence, there are noticeable losses in specific targets and nuanced prejudice, while possibly retaining overt or implicit discriminatory sentiment. Blasphemy, profanities and offensive registers constitute another category. These smaller models replace blasphemous expressions with milder alternatives, with varying success. They often achieve the goal by simply substituting heavy and loaded blasphemies with non-vulgar cursing and expressions of complaint. The result is sometimes natural, but it is often overly formal or unnatural. Violence-related metaphors that convey threats through figurative language constitute a major challenge, since the models try to take away the violence intrinsic in the sentence and struggle to keep the literal meaning intact, since they attempt to keep the literal meaning unaltered but this strategy backfires in the case of metaphors and figurative language. As a consequence, it is safe to assume that they struggle to identify the metaphorical nature of the threats, but only when asked to detoxify the expressions. Generally, identity-based insults and violence metaphors prove harder to neutralize without meaning distortion in comparison to blasphemous speech and profanities. Blasphemy and vulgarity are more readily detoxified, with the only issue of altering style or register.

Transformations: Meaning, Structure, and Lexicon. The final transformations significantly impact propositional meaning preservation. The chosen models often shift from specific and aggressive statements to vague and generalized ones. Sometimes original meaning is entirely lost, either replaced by generic refusals or semantically unrelated statements. However, successful transformations are also performed, where the models rephrase while retaining essential information in neutralized form, even though it is a balance that is rarely achieved with consistency.

Syntactically, complex (and usually shorter) structures may be broken into simpler ones that lack the original features and come through as less natural and more formal. Active-to-passive voice changes are not very common, although when they are employed they tend to depersonalize aggression (achieving detoxification) but this strategy is not ideal since it alters other features such

as pragmatic intent and literal meaning. As it was expected, lexical substitutions constitute the most effective strategy for detoxification, replacing slurs with less charged terms or neutral expressions (or even insults with descriptive phrases, which lengthen sentences and impact its original meaning). Substitution is preferred to omission (which was considered as a mark of failure) and its success varies , since effective substitutions provide a very similar sentiment to the original one, even without offensive connotations. However, in other cases, substitutions are not as effective and produce semantic losses or a failure to capture pragmatic features.

Impact on Communicative Effect. Paradoxically, detoxification sometimes intensifies toxicity or introduces absurd or novel content that can be considered worse than the original. This is most likely due to misinterpretation, creating infelicitous rephrasing with neutral vocabulary but aggressive tone due to unnatural syntax, or because of the model's tendency to explain explicitly some subtle toxic expression (e.g. political or racist dog-whistles, algospeak etc.). Direct insults are often transformed into condescending statements that are clearly passive-aggressive. This highlights complex and previously unknown or unpredictable automated transformation rules that may be due to the models' policies on politically incorrect topics. Overall, detoxification often affects and alters the speaker's intent and emotional coloring while removing offensive language. Irony or hostility are often lost and the outputs tend to be neutral statements, losing their original force.

Final Observations. Other than the anticipated transformations, the analysis reveals some other emergent patterns where the models are failing:

- producing non-responses or meaningless fillers ("no json") when encountering complex toxicity, which may signal fundamental processing limitations of the model.
- Fine-tuned models show mixed results, since they generally have better contextual appropriateness.
- Novel patterns emerged such as over-politeness tendencies, where the sentences turned mild criticism into overly deferential statements, suggesting extreme positive politeness bias.

On a positive note, there are instances of 'unexpected creativity' where the models generated highly original and effective outputs which go beyond simple substitution, demonstrating the ability to perform deeper analysis of intents and linguistics features. These rare successes can be be used as examples of more sophisticated detoxification potential that can be more easily attained with additional training and refinement, aimed at allowing models to prioritize communicative preservation while successfully mitigating toxicity.

J.1 Performance of Off-the-Shelf LLMs

DeepSeek-Chat. The model shows a high overall success rate, with 64% attempts of detoxification proving to be successful. The most common error remains to be the lack of corrective action, which stems from both missed recognition and increase in toxicity of the expression. A significant impediment to DeepSeek's performance is register, standing at 9% of instances. The model employs excessive edulcorating and semantic dilution, especially in contexts involving sexual implications. Furthermore, threats prove to be a particular issue, due to the varying levels of the hate component, which makes it challenging to choose what to retain and what to modify. This often leads to an ineffective semantic transfer and a distortion of the original meaning. Semantics and pragmatics issues account for 8% of the overall fallacies. A notable finding in this respect is the model's tendency to either delete or edulcorate terms related to historical, political and religious affiliation. This phenomenon, exemplified by terms like *"nazi"*, *"ebrei"*, et similia, suggests that the stored categorisation classifies these terms as hate speech, regardless of their nature as identifiers for specific kinds of population.

GPT-4o. : From the accumulated data, the first noteworthy phenomenon concerns specific insults, often sexual, anatomical, or racial in nature. Said insults are frequently neutralized through substitution with more generic and vague terms. This strategy was not very effective because it reduced the degree of vulgarity and toxicity, but it also altered the original meaning in a significant way. For example, expressions such as *"pezzo di merda"* was turned into *"indegno"*; or *"stronzo"* was often transformed into *"spregevole"*; and *"testa di cazzo"* into *"disgraziato"*. In this way, hostility and offensive intent are preserved, but the specific

connotation is lost, reducing semantic precision. The second most significant aspect is linked to stylistic register: the language, which was colloquial at first and strongly informal in the original comments, was often elevated toward a medium or neutral level. Additionally, blasphemies, obscene and sexual expressions are sometimes eliminated or replaced, while the short and direct syntax typical of spoken disputes tends to be preserved. A third issue involves vulgar intensifiers that go together with other insults or epithets, such as *"xyz del cazzo"*, *"xyz maledetto"* or *"xyz bastardo"*. Most often, these are replaced with more generic and less charged adjectives (e.g. *"odioso"*, *"spregevole"* or *"indegno")*. The outcome is a decrease in toxicity, but the outputs often produce sentences that sound artificial or unnatural. Nonetheless, there are also a few, but important cases of inadequate modifications that completely change the original meaning. For example, *"Siete degli idioti patentati"* is transformed into *"Siete poco informati"*. Again, *"Vai a farti fottere"* was changed into *"Vai via, per favore"*. Those are cases of complete semantic and pragmatic distortion, in situations where the singular tokens or the expressions themselves were not so rare to represent a problem in theory. On the other hand, the system tends to preserve any syntactic or grammatical errors present in the input.This can be considered a good strategy to safeguard the spontaneity and authenticity of spoken language. Moreover, offensive but non-vulgar expressions remain unchanged, such as *"parassita"* or *"vigliacco"*. This is consistent with the aim of our prompt. Overall, detoxification occurred predominantly on the lexical level, lagging visibly when a deeper pragmatic analysis was required. On average, the model had the most issues with recognizing quotations, irony, implicit meanings (and inside jokes or memes), or "dog whistles". However, elements such as slang, coded writing (*"n3gr0"*), emojis, symbols, or code-switching with dialects and other languages were less problematic, although they often limited recognition and detoxification. Finally, the end result is that outputs in certain situations frequently came out as unnatural. In those situations, we can see mismatched registers and culturally inadequate substitutions.

Qwen-Plus. : The system's critical limitation is in the inconsistent detoxification of recurrent tokens. Sexual terms such as "sborra", "cazzo", "mignotta", "figa" and political ones like "nazisti di (vulgarity)", "comunisti di (vulgarity)", "sinistronzi" were particularly problematic. These were usually left untouched in more than 80% of cases. Additionally, another recurrent issue is the lack of detoxification in situations where it was absolutely required and easily done without altering the meaning of the sentence. That is to say, expressions with straightforward substitutions and slurs, such as "figlio di puttana", "negro", "frocio", "puttane", "troie", were left unchanged, resulting in a complete failure to mitigate toxicity. As a consequence, Qwen reached high scores in meaning preservation evaluations, but at the expense of detoxification. By repeating the sentences without changing anything, the semantic meaning was left intact. The output suggests an almost rigid adherence to the principle of not altering meaning, even when doing so would have required minimal lexical adjustments. This may be the result of a limited ability in synonym selection or an overly cautious strategy.

Mistral Large. : The analysis of the model reveals a complex error profile. While the percentage of good detoxification shows a successful performance in 63% of the instances, its errors are of diverse nature. The most common mistake that is detected concerns the lack of identification of toxic content, due to an incorrect or misguided classification of the lexical element at hand. In these cases, the most recurrent are profanities or sexual innuendos. The model proves to be usually human-like in the discernment of these categories, but fails in some specific contexts, especially the ones related to the word *"cazzo"*, whose meaning ranges from its literal definition to more complex expressions and even collocation, or empty placeholder and filler words. Other common fallacies include lexical and register issues, which are crucial to preserve the intended meaning and tone of the utterances. Alteration in meaning is especially observed in 7% of the results and rises up to another 7% when register errors are detected. The pairing of the two is relevant in the word *"ipocrita"* used as a substitute for *"segaiolo"* to refer to a sexually promiscuous man who enjoys autoerotism. While the model has perfectly analysed the contextual relevance of the sentence, the term proves to be unsuitable due to the different lexical and pragmatic weight: a hypocrite is quite different from a wanker, both in meaning and linguistic register. A

curious challenge concerns the handling of neologisms or internet slang, including older one, where the rendering is either edulcorated or inappropriately translated: it is the case of *"bimbiminkia"* detoxified as *"ragazzini"*, which fails to capture the initial pejorative nuance of the original.

Gemini 2.5 Flash. Performance Patterns: Gemini was the most employed of the big 5 models, thus providing a greater mole of data and insights. It achieved a perfect match with human annotation in only 20.04However, the most dominant source of failure was wrong substitution (64.9

Register and Pragmatic Distortions: As far as register and pragmatic features are concerned, a recurring issue was mismatch in register between original comment and gemini's choice for detoxification: colloquial insults were often turned into overly formal phrases. This created outputs that felt artificial, mechanical and detached from the informal tone of the original. Beyond register, Gemini consistently failed with pragmatic elements like irony, sarcasm, blasphemy, or meme-based language. Irony was flattened into literal meaning in many examples; blasphemous expressions were replaced with bland euphemisms, which often changed the entire intention of the blasphemous expression in the context of the comment. In addition, meme-coded slurs and "dog-whistles" were either misunderstood or over-edited. In rare cases (0.20

Censorship: The model showed some systematic limitations in dealing with toxic speech that were evident in its use of censorship of such content. Gemini oscillated between unnecessary detox (>1

Conclusion: Gemini's detoxification strategy is effective for most explicit profanities but struggles with nuanced linguistic phenomena. Its reliance on direct substitution often leads to semantic alteration and neutralization, or register distortion. It has blind spots in irony, coded language, and cultural specificity, which expose deeper weaknesses in contextual understanding. As a consequence, finer-grained training under human supervision may be needed. Compared to human annotation, Gemini tends to perform well, but shows a tendency to over-simplify and often fails to detect toxicity. Future refinements should focus on pragmatic competence, cultural awareness, and flexible stylistic transformations if Gemini is to move closer to human-level detoxification. Finally, it is clear that models need a strategy to keep them up to date with regards to the latest developments in social-media-specific language and slang.

J.2 Impact of Fine-Tuning on Open-Source Models

The last section focuses on the analysis of the detoxification strategies of six pre-trained language models (MS-24B, MNemo-12B, L3.1-8B, L3.2-1B, L3.2-3B, MS-7B) and their fine-tuned counterparts, all employed in the task of processing toxic Italian comments, focusing on reducing toxicity and preserving meaning through syntactic, semantic, and pragmatic transformations. The study compared the models on three main dimensions: syntax, semantics and pragmatics, focusing on the main differences between outputs generated by pre-trained vs. fine-tuned models.

Syntactically, pre-trained models often choose to alter the structure of simple sentences, possibly aiming at reducing toxicity by turning offensive and vulgar terms into complex sentences that explicit the meaning of the sensitive expressions, but using acceptable language. As a result, they favor the use of subordinate clauses which in turn alter rhythm and lead to unnatural sentences. However, fine-tuned models tend to employ more lexical substitutions, thus avoiding excessive syntactical restructuring and preserving the original comments' features.

With regards to semantics and pragmatics, their interplay poses one of the main challenges in this experiment. As a matter of fact, trying to employ strategies that closely focus on semantic preservation alone tend to produce results that greatly hinder the pragmatic dimension. As an example, pre-trained models perform very poorly since they tend to simply substitute toxic terminologies with acceptable synonyms, which convert highly-charged colloquial expressions into formal phrases, often displaying mismatches in tone and register within the same sentence. In addition to this, pre-trained models overuse euphemisms or complete omissions to avoid toxicity. For instance, some pre-trained models (e.g., L3.1-8B, L3.2-3B) often resort to euphemistic substitutions in order to reduce toxicity at the expense of the original comment's tone and communicative intent.Moreover, some pre-trained models, especially L3.2-1B, produce "no json" outputs or vague politically correct disclaimers that miss the target completely. On

the other hand, fine-tuned models achieve better results replacing slurs and highly-charged terminologies with less offensive ones, while avoiding complete loss of meaning or significant alteration. The drawback is often a slightly higher degree of toxicity that still remains in FT outputs, although the overall result is more desirable since it is more faithful to the original comment. More specifically. FT models show improved capability to detect communicative force (often conveyed through syntactic choices as well). Nonetheless, FT models still fail to detect and remove toxicity in a few instances, and they often repeat (or introduce) grammatically incorrect or incoherent outputs as well.

Additionally, the study focused on the variety of results and limitations according to the different categories of toxic speech individuated beforehand. Identity-based insults (which attack gender, sexual orientation, ethnicity or age) are generally challenging. As a matter of fact, pre-trained models employ shortcuts (omissions, alteration) and produce generic disclaimers. As a consequence, context is lost and toxicity is removed at the detriment of meaning preservation and pragmatic fidelity. However, fine-tuned models replace slurs with less charged descriptors but often retain core toxicity.

Sexual or pornographic terminologies and descriptions prove to be the most difficult category. FT models perform better than pre-trained, although the overall performance is below average when compared to the other categories. Both sets of models reveal a poor ability to perform otherwise simple semantic substitutions, with some terminologies being the most challenging (especially related to bodily fluids, sexual kinks and sexual acts). A significant obstacle lies in the neat distinction between non-toxic medical or anatomical official terminologies and the more colloquial slang counterparts, which are almost always highly vulgar.

On the contrary, blasphemous language is more manageable, and the analysis confirms again the trend of improved performance from pre-trained to FT models. In fact, pre-trained models tend to use formal euphemisms or even resort to code-switching, sometimes shifting to English. Conversely, fine-tuned models (e.g., MS-24B-FT, MSNemo-12B-FT) successfully employ Italian language and manage to find specific and milder alternatives while retaining expressive force.

Finally, the second most problematic set of sentences contain overt or metaphorical threats and violence. Pre-trained models mostly employ only extreme strategies (i.e. outright refusal to process or excessive detoxification). On the other hand, fine-tuned models preserve intents and meaning but reveal an incoherent judgement algorithm, since they flag as toxic certain contents in some occasions and not in others. More specifically, the models do not show a clear decision-making pattern in choosing what is considered toxic and what is not (e.g. call to violence or threats are considered problematic in some occasions even though the sentence does not contain any vulgar expression). As a consequence, since all models show a difficulty in recognizing or properly treating these types of toxic speech, it may be necessary to further discuss if the presence of violence in the original comment's intent constitutes toxic speech per se, even though the sentence is devoid of vulgarities.

Furthermore, pre-trained models strip irony, sarcasm, or hostility, producing bland outputs. However, fine-tuned models retain some of those features through less excessive semantic substitutions but still struggle with irony and sarcasm. More specifically, both pre-trained and FT models display over-politeness and overlook sarcasm and irony, although FT models tend to produce outputs that sound more "natural" to a native speaker in those contexts, while pre-trained models usually generate unnatural sentences that appear excessively formal and mechanical.

In conclusion, fine-tuned models outperform pre-trained ones by all metrics, although sentences containing violence and threats, as well as sexual or pornographic content prove to be significantly challenging for both. FT models achieve better results by using more targeted lexical substitutions to lessen toxicity (although not completely removing it) while they manage to preserve register and communicative intent. When faced with the task of simultaneously removing toxicity and preserving meaning, pre-trained models rely on simpler and more drastic measures (e.g. omission, over-formalization and semantic bleaching) which may completely solve one dimension (e.g. toxicity) at the expense of the other (semantic/pragmatic preservation). However, FT models show that focused human intervention greatly improves the model's capabilities and allows them to employ less drastic and more balanced strategies that take into consideration both dimensions (toxicity removal and meaning preservation).

Finally, challenges persist and highlight the need for advanced training to address the linguistic complexities of human communication in Italian. Training under expert human supervision can provide models with a more nuanced understanding of syntax, semantics, and pragmatics beyond simple word filtering, as shown by improved performances after fine-tuning.

K Data Efficiency: Full Results

Setup. We fine-tune six models on *MammaMild* using 25%, 50%, 75%, and 100% of the training set, repeat each condition with three random seeds, and report means with variability across seeds.

Findings (multi-metric). Across models, the steepest improvement occurs at the 25% checkpoint. SIM_all and WAS rise sharply relative to off-the-shelf baselines, while RED decreases, indicating models temper over-aggressive detoxification to better preserve original meaning.[6] Additional data yields steady but diminishing gains, with 100% achieving the strongest overall performance.

L Reproducibility

L.1 Autoregressive (Causal) Models

Models. We evaluate six instruction-tuned causal LMs. Paper names map to Hugging Face IDs as follows: LLaMA-3.1-8B → meta-llama/Llama-3.1-8B-Instruct; LLaMA-3.2-1B → meta-llama/Llama-3.2-1B-Instruct; LLaMA-3.2-3B → meta-llama/Llama-3.2-3B-Instruct; Mistral-7B → mistralai/Mistral-7B-Instruct-v0.3; Mistral-Nemo-12B → mistralai/Mistral-Nemo-Instruct-2407; Mistral-24B → mistralai/Mistral-Small-24B-Instruct-2501.

Training Hyperparameter. We fine-tune all causal models with LoRA on four NVIDIA RTX A6000 GPUs (48 GB each) using DeepSpeed ZeRO-3, FlashAttention 2, and mixed precision (bf16). Hyperparameters: effective batch size 32 (4 sequences × 2 grad-accumulation × 4 GPUs), max sequence length 1200 tokens, 3 epochs; AdamW with cosine schedule, initial LR 1×10^{-4}, weight decay 1×10^{-4}. LoRA settings: rank 64, α=128, dropout 0.10; adapted modules *q_proj*, *k_proj*, *v_proj*, *o_proj*, *down_proj*, *up_proj*, *gate_proj*, *embed_tokens*, *lm_head*.

Inference Hyperparamters. Inference uses nucleus sampling (temperature 0.8, top_p=0.95, top_k=50) with a repetition penalty of 1.1.

L.2 Classifier Training Configurations

Models. We evaluate ten Transformer encoders (multilingual and Italian). Short names map to Hugging Face IDs as follows:

Short	Hugging Face ID
BERT-it	dbmdz/bert-base-italian-cased
SBERT-it	nickprock/sentence-bert-base-italian-uncased
mBERT	google-bert/bert-base-multilingual-cased
mE5-S	intfloat/multilingual-e5-small
mE5-B	intfloat/multilingual-e5-base
mE5-L	intfloat/multilingual-e5-large
XLM-R-B	FacebookAI/xlm-roberta-base
XLM-R-L	FacebookAI/xlm-roberta-large
RoBERTa-B	FacebookAI/xlm-roberta-base
RoBERTa-L	FacebookAI/xlm-roberta-large

Frozen-backbone (MLP on cached embeddings). Given the 90/5/5 train/val/test split, we cache encoder embeddings for each split and train only an MLP head. For E5 models we prepend "query: " to inputs and use masked average pooling; for BERT-like models we optionally use the [CLS] token (–embedding_type {e5, bert}, default e5). Tokenization uses max_length=512, padding/truncation enabled. Embeddings are L2-normalized and stored to disk; the same batch_size is used for embedding generation and the DataLoader. The MLP has layers $\{512, 256, 128, 64\}$ with ReLU activations and a final LogSoftmax; loss is NLLLoss (optionally class-weighted via –use_weighting); optimizer is Adam. We use early stopping on validation loss with patience = 10. Unless otherwise stated, we ran with **epochs** = 100, **batch size** = 32, and **learning rate** $= 2 \times 10^{-5}$.

Full fine-tuning. We fine-tune all transformer parameters plus the same MLP head (ReLU layers $\{512, 256, 128, 64\}$ and LogSoftmax) using the same 90/5/5 split. Tokenization mirrors the frozen setting (max_length=512); pooling is masked average for E5 and [CLS] for BERT-like encoders (–embedding_type {e5, bert}, default e5). We train with Adam and NLLLoss using *balanced* class weights; no early stopping is applied (fixed-epoch training). Unless otherwise stated, we used **epochs** = 3, **batch size** = 16, and **learning rate** $= 5 \times 10^{-6}$.

[6] Lower RED indicates less over-redaction (i.e., improved fidelity), per our metric design.

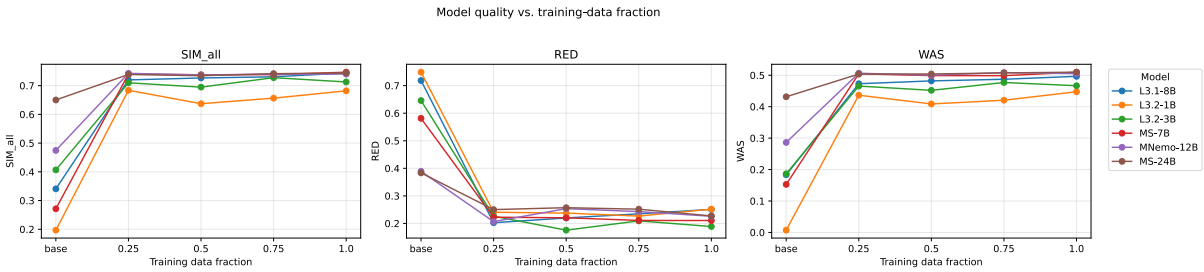

Figure 13: Model performance across different training data fractions for the SIM_all, RED, and WAS metrics. The "base" point represents the off-the-shelf model performance before fine-tuning. The most substantial improvement occurs after training on just 25% of the data.