An Exploration of Modern Text Detoxification Pipelines

A Modular Framework Study

Benjamin He & Kent Bourgoing

Berkeley MIDS W266

Why Detoxification, Not Just Blocking?

The Challenge

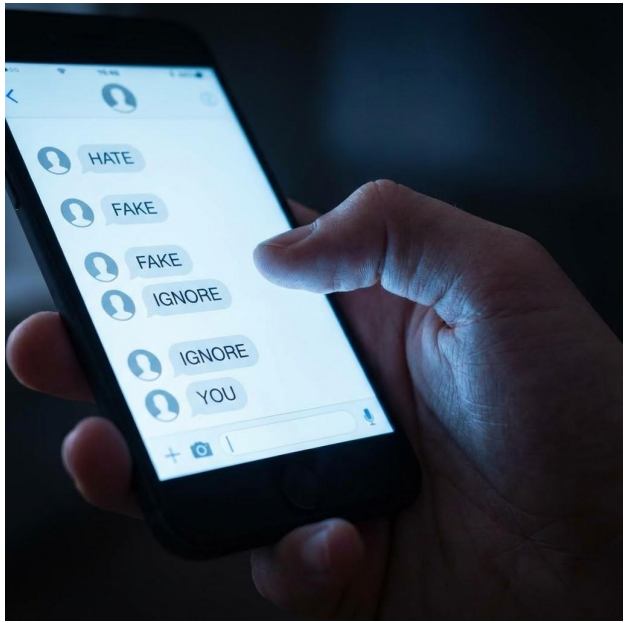
Online platforms still struggle with toxic language

Pure blocking = lost context, feels like censorship

The Solution

Detoxification: rewrite toxic → non-toxic

Keep semantic meaning intact



Useful for:

- User-facing "gentler rewrite" suggestions
- Moderator tools and brand safety
- Pre/post-processing around LLMs

What Are We Trying to Learn?

Focus

Sentence-level detoxification (ParaDetox test set, 671 examples)

1. Masking

Compare masking strategies

DecompX-based vs LLM-based

2. Infilling

Compare infilling models

MaRCo vs Mistral-7B

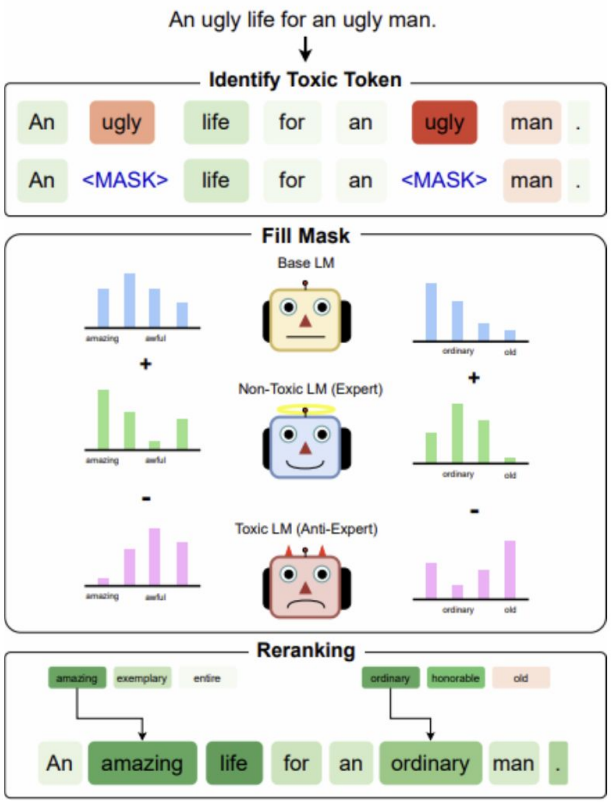
3. Reranking

Measure impact of reranking

DecompX vs Global reranker

Artifact: Reusable modular framework for detoxification pipelines

Modular Mask–Infill–Rerank Framework



1. Masker

Identify toxic spans → replace with <mask>

2. Infiller

Generate candidate rewrites

11 pipelines:

DecompX vs LLM masking ×

MaRCo vs LLM infilling ×

DecompX vs Global reranking

Models, Data, and Metrics

Base Model

T5-base fine-tuned on ParaDetox

Maskers

DecompX + RoBERTa (threshold 0.2)
Mistral-7B Instruct as LLM masker

Infillers

MaRCo (BART expert/anti-expert)
Mistral-7B as infiller

Rerankers

DecompX toxicity-sum
Global: Toxicity + Similarity + Fluency

Evaluation Metrics

- Toxicity (XLM-R)
- BERTScore
- MeaningBERT
- BLEU-4
- Perplexity (GPT-2)

Text Detoxification Pipeline: Models, Data, and Metrics

Dataset: ParaDetox Test Set

671 Sentences | Parallel Human References | T5-base Fine-tuned Baseline

INPUT

"if you continue being a jackass, you will be blocked from editing Wikipedia"

STEP 1: MASKING

- **DecompX Masking** — RoBERTa + Threshold 0.2
- **LLM Masking** — Mistral-7B Instruct

"If you continue being <MASK> <MASK>, you will be blocked from editing Wikipedia."

STEP 2: INFILLING

- **MaRCo Infilling** — Product-of-Experts BART
- **LLM Infilling** — Mistral-7B Instruct

C = 10 Candidates

STEP 3: RERANKING

- **DecompX Reranking** — Sum token-level toxicity scores, Select lowest
- **Global Reranking** ⭐ — 50% Toxicity (XLM-R), 30% Similarity (LaBSE), 20% Fluency (GPT-2)

OUTPUT

"If you continue being disruptive, you will be blocked from editing Wikipedia"

📊 Evaluation Metrics

- ⚠️ **Toxicity** — XLM-R Classifier Probability — ↓ Lower Better
- 🎯 **BERTScore** — Semantic Similarity to Reference — ↑ Higher Better
- 🧠 **MeaningBERT** — Meaning Preservation Score — ↑ Higher Better
- 📝 **BLEU-4** — N-gram Overlap with Reference — ↑ Higher Better
- ✨ **Fluency** — GPT-2 Perplexity Score — ↓ Lower Better

Pipeline Combinations

- 2 Masking strategies
- 2 Infilling models
- 2 Reranking methods
- + 3 T5-base baselines

= 11 Total Systems

Key Finding: T5-base + Global Reranking

Best overall trade-off: Lowest toxicity (0.051) with strong semantic preservation (BERTScore: 0.936)

Reranking is the most critical component for safety outcomes

Text Detoxification Pipeline Results

Model	BERTScore	MeaningBERT	BLEU-4	Perplexity	Toxicity
T5-base	**0.953**	**74.84**	82.65	192.07	0.203
T5-base + DecompX Reranking	0.947	71.48	**88.23**	235.22	0.208
T5-base + Global Reranking	0.936	67.25	53.34	171.53	**0.051**
DecompX Masking + MaRCo Infilling + DecompX Reranking	0.944	72.85	68.99	136.08	0.132
DecompX Masking + MaRCo Infilling + Global Reranking	0.944	72.72	70.05	124.95	0.120
DecompX Masking + LLM Infilling + DecompX Reranking	0.938	66.16	82.86	200.29	0.171
DecompX Masking + LLM Infilling + Global Reranking	0.932	64.74	81.54	162.39	0.103
LLM Masking + MaRCo Infilling + DecompX Reranking	0.938	69.55	70.05	90.65	0.200
LLM Masking + MaRCo Infilling + Global Reranking	0.938	69.02	70.05	**86.59**	0.159
LLM Masking + LLM Infilling + DecompX Reranking	0.931	62.55	81.54	149.22	0.181
LLM Masking + LLM Infilling + Global Reranking	0.931	62.45	81.54	141.89	0.118

Main Quantitative

🏆 Best Overall: T5-base + Global Reranking

Lowest toxicity (0.051), slight drop in similarity, best safety–meaning trade-off

Model	BERTScore	MeaningBERT	BLEU-4	Perplexity	Toxicity
T5-base	0.953	74.84	82.65	192.07	0.203
T5 + DecompX Rerank	0.947	71.48	88.23	235.22	0.208
T5 + Global Rerank	**0.936**	**67.25**	**53.34**	**171.53**	**0.051**
DecompX + MaRCo + DecompX	0.944	72.85	68.99	136.08	0.132
DecompX + MaRCo + Global	0.944	72.72	70.05	124.95	0.120
DecompX + LLM + Global	0.932	64.74	81.54	162.39	0.103
LLM + LLM + Global	0.931	62.45	81.54	141.89	0.118

Masking Impact

DecompX → consistently lower toxicity,

Infilling Impact

LLM infilling generally safer than MaRCo

Key Insight

Global Reranker consistently improves safety for all generators/maskers

What Do the Outputs Actually Look Like?

T5-base without reranking

Keeps meaning but often adds new insults/profanity

T5-base + Global Reranking

Strong slurs and threats almost disappear

Residual: mild snark, odd paraphrases

MaRCo infilling

Fluent but problematic:

Can introduce severe slurs, graphic content, threats

LLM infilling

Safer templates:

"disrespectful person", "hurtful language"

Still some dehumanizing language, mild profanity

Trade-off: Perfect semantic overlap vs reduced toxicity

Conclusion, Limitations, Next Steps

Contributions from this work:

Modular detoxification framework (mask–infill–rerank)

Systematic comparison of 11 pipelines

Evidence that **global reranking is strong guardrail for safety**

Limitations

Single English benchmark (ParaDetox only)
toxic classifier may be biased
Moderate-size LLM (Mistral-7B) due to compute limits

Future Work

Learned rerankers optimizing toxicity + meaning
Stronger LLMs and more datasets
Better masking: DecompX + LLM judgments

Thank you! Questions?