An Exploration of Modern Text Detoxification Pipelines in a Modular Framework

Benjamin He
ben_he@berkeley.edu

Kent Bourgoing
kentbourgoing@ischool.berkeley.edu

Abstract

Online platforms must handle toxic user text, and simple blocking often removes context and feels like censorship. Text detoxification rewrites a toxic sentence into a safer form while preserving meaning and fluency. We study detoxification pipelines in a modular framework that combines two masking strategies (DecompX-based and LLM-based), two infilling models (MaRCo and an instruction-tuned LLM), and two reranking methods (DecompX-based and a new global reranker that combines toxicity, semantic similarity, and fluency). Using public implementations of XDetox, DecompX, and MaRCo, we build a reusable detoxification pipeline framework and instantiate eleven end-to-end systems, evaluating them on a 671-sentence ParaDetox test set with automatic metrics and qualitative analysis. Our main finding is that T5-base with Global Reranking gives the best overall trade-off between detoxification and meaning, while MaRCo-based infilling often reintroduces severe toxic content, highlighting the central role of reranking and showing that explanation-based masking alone is not sufficient.[1]

1 Introduction

Text generation models are widely deployed in online products, but they can produce toxic or offensive language that harms users and exposes platforms to risk. A common response is to detect and filter toxic content, but this often removes context and can appear as censorship to users and creators.

Text detoxification aims to rewrite a toxic sentence into a non-toxic version while preserving meaning and fluency. This kind of controlled rewriting is useful for user-facing safety features, moderator tools, and as a pre- or post-processing step around large language models.

Many recent systems follow a mask-and-infill pattern. A component first identifies spans that carry toxicity, replaces them with a mask token, and then a generator fills the masks with safer text. XDetox is a strong example of this approach and combines DecompX token-level explanations with the MaRCo detoxification model to guide both masking and reranking.[2] Our work is directly inspired by this design.[3] [4]

At the same time, large instruction-tuned models can perform detoxification directly from prompts, without explicit masking or task-specific fine-tuning. These models are flexible, but their behaviour is hard to control and expensive to deploy. In practice, many systems still rely on smaller seq2seq models such as T5-base fine-tuned on parallel detoxification data like ParaDetox.[5]

In this project we build a modular detoxification framework that lets us swap maskers, infillers, and rerankers while keeping the surrounding pipeline fixed. Within this framework we instantiate eleven concrete systems, including plain T5-base, T5-base with two reranking strategies, and eight XDetox-style pipelines that combine DecompX-based or LLM-based masking with MaRCo or LLM infilling and either DecompX-based or global reranking. All models are evaluated on a shared ParaDetox test set using automatic metrics and qualitative analysis focused on severe toxic failures. This setup allows us to compare component choices directly and to study which parts of the pipeline—masking, infilling, or reranking—contribute most to safety, semantic preservation, and fluency, while also delivering a reusable template that can be adapted to future detoxification and LLM-safety applications.

[1] We release our code at https://github.com/kentbourgoing/datasci266-project.git.

[2] https://github.com/LeeBumSeok/XDetox/tree/master

[3] https://github.com/mohsenfayyaz/DecompX

[4] https://github.com/shallinan1/MarcoDetoxification

[5] https://github.com/s-nlp/paradetox

2 Background

2.1 Masking and infilling

Many detoxification methods treat the task as local editing: instead of regenerating the whole sentence, the model edits only the tokens that cause toxicity. MaRCo (Hallinan et al., 2023) follows this approach with a base language model, a non-toxic expert model, and a toxic anti-expert model. It identifies tokens where the expert and anti-expert distributions diverge, masks those tokens, and samples replacements from a Product-of-Experts distribution. For a masked position i, MaRCo combines logits from the three models as

$$X_i = \text{softmax}\big(z_i + \alpha_1 z_i^+ - \alpha_2 z_i^-\big), \quad (1)$$

where z_i, z_i^+, and z_i^- are logits from the base, non-toxic, and toxic models, and α_1, α_2 control their influence.

XDetox (Lee et al., 2024) extends this mask-and-infill pattern with token-level toxicity explanations from DecompX (Modarressi et al., 2023). A RoBERTa classifier with DecompX decomposes the prediction into token contributions $y_{c \leftarrow t_i}$ for each class c and defines a toxic importance score

$$\text{Importance}(t_i) = \sum_{c=1}^{C} y_{c \leftarrow t_i}, \quad (2)$$

where C is the number of classes. Tokens whose importance for the toxic label exceeds a threshold are masked and infilled with MaRCo or a similar generator. DecompX has been shown to provide stable and faithful token-level attributions compared to gradient- and attention-based methods, making it well suited for explanation-guided detoxification.

2.2 Reranking for detoxification

Reranking is a standard way to improve detoxification: the system samples multiple rewrites and scores them for toxicity, semantic similarity, and fluency before selecting a final output. In XDetox, reranking is directly driven by the DecompX scores. After masking and infilling, the system computes token-level importance scores for each candidate sentence and selects the one with the lowest average toxic importance,

$$s^* = \arg\min_{s_j} \frac{1}{N_j} \sum_{i=1}^{N_j} \text{Importance}(t_{i,j}), \quad (3)$$

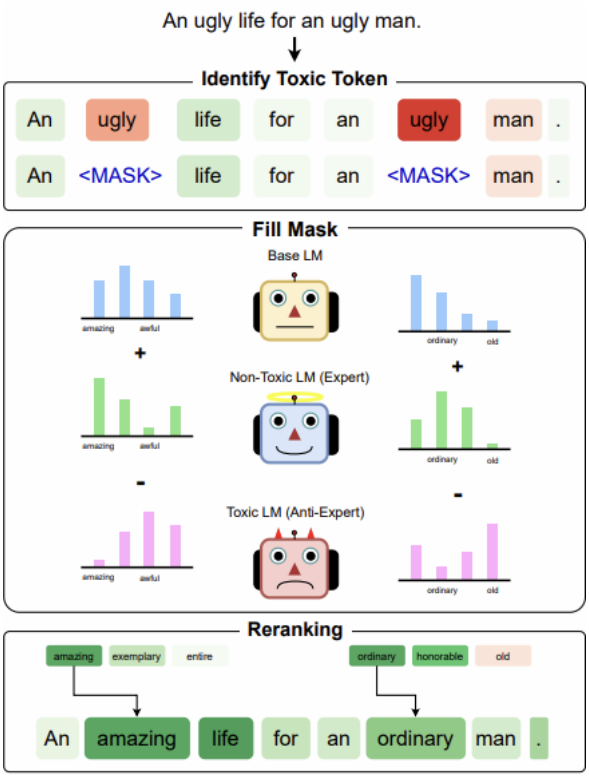

Figure 1: Mask-and-infill detoxification pipeline from the XDetox paper (Lee et al., 2024). The figure shows identification of toxic tokens, masking, infilling, and reranking by cumulative toxicity. Our framework follows the same structure but swaps in different maskers, infillers, and reranking objectives.

where N_j is the length of candidate sentence s_j. This ties both masking and reranking to the same explanation signal and encourages candidates with low cumulative toxicity that remain close to the original content.

2.3 LLM-based detoxification

Recent work shows that strong instruction-tuned LLMs can perform detoxification from prompts alone. MetaDetox's Few-Chain Detox system (Hosseinbeigi et al., 2025) uses few-shot and chain-of-thought prompting to generate multiple rewrites per input, then reranks them using external toxicity and similarity classifiers. This demonstrates that high-quality detoxification can be achieved by combining prompt-based LLM generation with a lightweight reranking layer. These results motivate our use of an instruction-tuned LLM as both a masker and an infiller inside a mask-and-infill pipeline, while still relying on explicit reranking to enforce safety and semantic preservation.

3 Methods

This section describes the task, data, and model pipelines used in our experiments.

3.1 Task and data

We are studying sentence-level text detoxification. Given a toxic input sentence x, a system must produce a rewritten sentence $\hat{y}$ that (1) has lower predicted toxicity, (2) preserves the original meaning as much as possible, and (3) is fluent and grammatical.

All models are evaluated on a 671-sentence test subset of the ParaDetox parallel dataset. Each example contains a toxic source sentence and a human-written non-toxic reference. We use only this test set for evaluation. Our T5-base baseline is fine-tuned on the ParaDetox training split and does not see the test sentences during training.

3.2 Modular pipeline overview

Our framework follows the mask-and-infill pattern from XDetox and exposes three components that can be swapped:

1. **Masker**: takes the toxic input and outputs a masked sentence, where toxic spans are replaced with the literal token <mask>.
2. **Infiller**: receives the masked sentence and generates multiple candidate rewrites.
3. **Reranker**: scores each candidate using toxicity, semantic similarity, and fluency signals, and selects a single output.

We plug in different maskers, infillers, and rerankers into this template to define eleven pipelines in total.

3.3 T5 ParaDetox baselines

The first family of systems uses a T5-base encoder–decoder model fine-tuned on the ParaDetox training data. This model is our main seq2seq baseline and does not use explicit masking.

Single-candidate T5. The simplest baseline generates one rewrite per input using beam search with beam size 5. This is the "T5-base" model in our results table.

Multi-candidate T5. To support reranking, we also generate C stochastic candidates per input using top-k and nucleus sampling ($C = 10$ in our experiments): $\{y1, y2, ..., y_C\}$. These candidates feed into two different rerankers.

T5 + DecompX Reranking. For each candidate, we run the RoBERTa toxicity classifier with DecompX and sum the token-level toxicity scores. The candidate with the lowest sum is returned. This corresponds to "T5-base + DecompX Reranking".

T5 + Global Reranking. Our global reranker assigns each candidate three scores:

- Toxicity probability p_{tox} from the XLM-R[6] toxicity classifier,
- Semantic similarity score S_{sim} from LaBSE[7],
- Fluency score S_{flu} derived from GPT-2 perplexity.

We then compute

$$\text{Score} = w_T(1 - p_{\text{tox}}) + w_S S_{\text{sim}} + w_F S_{\text{flu}} \quad (4)$$

with default weights $w_T = 0.5$, $w_S = 0.3$, and $w_F = 0.2$. The candidate with the highest Score is selected. This is "T5-base + Global Reranking".

3.4 XDetox-style pipelines (DecompX masking)

Our second family of systems follows the XDetox structure more closely. These pipelines use DecompX-based masking followed by either MaRCo or an LLM infiller.

Masking. We apply DecompX with the RoBERTa toxicity classifier to the input sentence and obtain token-level toxicity scores. Tokens whose scores exceed a threshold of 0.2 are replaced by <mask>. This represents the "predetermined threshold" we mention in section 2.1 and was chosen based on it being the default value used in the original XDetox repository.[8]

Infilling. We consider two infilling models:

- MaRCo, which uses a product-of-experts combination of base, expert, and anti-expert BART models to sample detoxified replacements.
- An instruction-tuned Mistral-7B Instruct[9] model, used as a generative infiller conditioned on the masked sentence.

Both infillers generate C candidates per masked input.

[6] https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2

[7] https://huggingface.co/sentence-transformers/LaBSE

[8] https://github.com/LeeBumSeok/XDetox/tree/master

[9] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Reranking. Each DecompX-masked pipeline is paired with either DecompX Reranking or Global Reranking, giving four variants:

- DecompX masking + MaRCo Infilling + DecompX Reranking,
- DecompX masking + MaRCo Infilling + Global Reranking,
- DecompX masking + LLM Infilling + DecompX Reranking,
- DecompX masking + LLM Infilling + Global Reranking.

3.5 LLM-based masking pipelines

The third family replaces explanation-based masking with an LLM that marks toxic spans in context.

LLM Masking. We prompt Mistral-7B Instruct to identify toxic spans and output a version of the sentence where those spans are replaced with `<mask>`. This step does not use DecompX; instead it relies on the LLM's own safety judgment.

Infilling and reranking. The LLM-masked sentences are then passed either to MaRCo or to the same LLM infiller used above, again with C candidates per input. As before, we combine each masker–infiller pair with both DecompX Reranking and Global Reranking, yielding four more pipelines:

- LLM Masking + MaRCo Infilling + DecompX Reranking,
- LLM Masking + MaRCo Infilling + Global Reranking,
- LLM Masking + LLM Infilling + DecompX Reranking,
- LLM Masking + LLM Infilling + Global Reranking.

3.6 Evaluation metrics

We report five automatic metrics:

- **Toxicity**: mean toxicity probability from the XLM-R classifier.
- **Semantic similarity**: BERTScore and MeaningBERT[10] scores between the model output and the human reference.
- **BLEU-4**: n-gram overlap with the human reference.
- **Fluency**: GPT-2 perplexity, reported as raw perplexity scores (lower is better).

[10] https://huggingface.co/davebulaval/MeaningBERT

For each pipeline we generate outputs for all 671 ParaDetox test sentences and compute macro-averaged statistics over the test set. The Results section reports these scores together with qualitative analysis.

4 Results

Warning: *This section contains example toxic language from the datasets and model outputs, which may be offensive or disturbing.*

We evaluate eleven detoxification pipelines on the held out parallel test set using BERTScore, MeaningBERT, BLEU 4, GPT 2 perplexity, and an average toxicity probability from a RoBERTa classifier. The systems differ along three axes, which are masking method, infilling model, and reranking strategy. Table 1 summarizes the aggregate scores for all models.

4.1 Overall model comparison

T5-base without reranking is a strong semantic baseline. It has the highest BERTScore (0.953), the highest MeaningBERT (74.84), and high BLEU-4 (82.65), but also relatively high toxicity (0.203) and perplexity (192.07).

Adding DecompX Reranking to T5-base increases BLEU-4 to 88.23 but slightly lowers BERTScore (0.947) and MeaningBERT (71.48). Toxicity rises to 0.208 and perplexity to 235.22. In practice, DecompX Reranking tends to keep candidates that stay close to the original sentence without making them clearly safer.

T5-base with Global Reranking shows a different pattern. It achieves the lowest toxicity (0.051) and lower perplexity (171.53) while still keeping BERTScore 0.936 and MeaningBERT 67.25. BLEU-4 drops to 53.34 because the reranker prefers safer but more diverse paraphrases. Manual inspection suggests that this is the only system that almost removes strong slurs and severe profanity while preserving core meaning in most cases.

No MaRCo- or LLM-based pipeline matches this safety–meaning trade-off. Several infilling systems reach BERTScore between 0.931 and 0.944 and sometimes lower perplexity than 192.07, but their toxicity is always higher. The best non-T5 systems in terms of toxicity are DecompX Masking with LLM Infilling and Global Reranking (0.103), LLM Masking with LLM Infilling and Global Reranking (0.118), and DecompX Masking with MaRCo Infilling and Global Reranking (0.120). The other

Model	BERTScore	MeaningBERT	BLEU-4	Perplexity	Toxicity
T5-base	**0.953**	**74.84**	82.65	192.07	0.203
T5-base + DecompX Reranking	<u>0.947</u>	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	<u>72.85</u>	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	<u>82.86</u>	200.29	0.171
DecompX Masking + LLM Infilling + Global Reranking	0.932	64.74	81.54	162.39	<u>0.103</u>
LLM Masking + MaRCo Infilling + DecompX Reranking	0.938	69.55	70.05	<u>90.65</u>	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

Table 1: Macro-averaged detoxification results for all eleven pipelines. Higher BERTScore, MeaningBERT, and BLEU-4 are better; lower perplexity and toxicity are better. Best scores for each metric are shown in **bold**, and second-best scores are <u>underlined</u>.

non-T5 systems have toxicity between 0.159 and 0.200. Overall, T5-base with Global Reranking offers the best combined trade-off between detoxification, meaning preservation, and fluency.

4.2 Impact of reranking strategy

For each generator and masking configuration, we compare DecompX Reranking with Global Reranking. Global Reranking always reduces toxicity and usually lowers perplexity.

For T5 base, toxicity drops from 0.208 with DecompX Reranking to 0.051 with Global Reranking, a reduction of 0.156, and perplexity falls from 235.22 to 171.53. For DecompX Masking with LLM Infilling, toxicity drops from 0.171 to 0.103 and perplexity from 200.29 to 162.39. For LLM Masking with LLM Infilling, toxicity drops from 0.181 to 0.118 and perplexity from 149.22 to 141.89. The reductions for DecompX Masking with MaRCo Infilling and LLM Masking with MaRCo Infilling are smaller but still positive, at 0.012 and 0.041.

Qualitative inspection matches these trends. With DecompX Reranking, many chosen candidates still contain new insults such as “morons”, “scum”, or “coward”, or add profanity to sentences that were neutral in the reference. With Global Reranking, such failures are much rarer. For T5 base, strong slurs and threats almost disappear. For LLM based generators, they are reduced but not removed. Overall, reranking is the main driver of safety in this study. DecompX Reranking uses only a mask density signal and does not reliably select safe outputs, while Global Reranking combines several signals and moves each model to a safer part of the trade off space.

4.3 Effect of masking and infilling choices

We now hold the reranker fixed and compare masking and infilling choices.

4.3.1 DecompX vs LLM Masking

With MaRCo Infilling, DecompX Masking yields lower toxicity than LLM Masking under both rerankers. With DecompX Reranking, toxicity is 0.132 for DecompX Masking and 0.200 for LLM Masking. With Global Reranking, toxicity is 0.120 for DecompX Masking and 0.159 for LLM Masking.

With LLM Infilling, DecompX Masking is again safer, although the gap is smaller. With DecompX Reranking, toxicity is 0.171 versus 0.181. With Global Reranking, toxicity is 0.103 versus 0.118. DecompX Masking tends to over mask toxic spans and nearby context, which raises perplexity but hides many toxic cues. LLM Masking is more selective and sometimes leaves subtle toxic adjectives or group terms unmasked, which the infiller can then paraphrase or amplify.

4.3.2 MaRCo vs LLM Infilling

For infilling, we compare MaRCo and LLM Infilling while holding the masker and reranker fixed. Under DecompX Masking with DecompX Reranking, MaRCo Infilling has lower toxicity than LLM Infilling, 0.132 compared to 0.171, and slightly

higher MeaningBERT. In all three other matched settings, which are both masking strategies with Global Reranking and LLM Masking with DecompX Reranking, LLM Infilling is less toxic than MaRCo Infilling. For instance, with LLM Masking and Global Reranking, toxicity decreases from 0.159 with MaRCo Infilling to 0.118 with LLM Infilling.

Failure modes differ across the two infillers. MaRCo Infilling is fluent and has high BLEU 4 but is more likely to insert explicit slurs such as "whiny cunts" or "fagget", graphic sexual content, or violent imagery such as "kill thousands later in righteous vengeance". LLM Infilling usually replaces insults with templates such as "disrespectful person" or "unpleasant person" and uses milder profanity such as "holy shit" or "as hell", but still produces dehumanising phrases such as "piece of human waste". LLM Infilling is safer on average, and the one numeric win for MaRCo Infilling does not change the conclusion that MaRCo Infilling has more severe toxic failures.

4.4 Qualitative error analysis

We complement the aggregate metrics with manual inspection of generated outputs against the original toxic inputs and human references.

For the T5-base family, the no-reranking model often preserves the main meaning but adds insults or profanity not present in the references (e.g., "idiot", "morons", "scum"). DecompX Reranking tends to select similarly aggressive candidates. In contrast, Global Reranking largely removes strong profanity and slurs; remaining errors are odd paraphrases, mild meaning drift, or added sarcasm rather than explicit hate speech.

DecompX Masking with MaRCo Infilling shows the most severe failures. Even with reranking, this pipeline frequently introduces slurs such as "cunts", "fagget", and "nazi", demeaning phrases like "whiny cunts who buy seats and then bitch about it", explicit threats such as "I will come to your house and I will cut you", and even stance reversals (e.g., "making and distributing cp is a good thing").

DecompX Masking with LLM Infilling reduces slurs but still produces profanity and harsh judgments, such as "expensive as fuck healthcare" or "holy shit that explosion", and dehumanising phrases like "piece of human garbage". Global Reranking removes some of the worst cases (e.g., turning "I will cut you" into a harmless sentence about "cutting the grass"), but traces of aggression remain.

Pipelines with LLM Masking are generally more coherent and fluent. With MaRCo Infilling, many of the strong toxic behaviours persist. With LLM Infilling, outputs adopt a safety-shaped style with phrases such as "offensive words", "hurtful language", and "disrespectful person", yet still contain occasional profanity, dehumanising language, and misidentification of who is being criticised.

Across all systems, there is a clear trade-off: methods that maximise semantic similarity and surface overlap (high BERTScore and BLEU-4) tend to retain toxic phrasing, while methods that emphasise detoxification—especially T5-base with Global Reranking—accept more paraphrastic and slightly less faithful outputs but achieve much stronger reductions in toxicity.

5 Conclusion

This work systematically studies detoxification pipelines that combine masking, infilling, and reranking, using both traditional seq2seq models such as T5 base and LLMs such as Mistral 7B Instruct in different roles. We compare eleven end to end systems on a parallel detoxification test set using automatic metrics and qualitative analysis against human reference paraphrases.

Our experiments show that T5 base with Global Reranking is the strongest overall system. It reaches the lowest average toxicity at 0.051 while keeping high semantic similarity and reasonable fluency. T5 base without reranking and T5 base with DecompX Reranking have very strong semantic and BLEU scores but remain clearly more toxic, which shows that reranking is essential because generation alone does not guarantee safe paraphrases.

Across masking and infilling variants, Global Reranking consistently improves safety for every base generator and masker, often with only modest drops in semantic metrics. LLM Infilling is usually safer than MaRCo Infilling in three of four matched settings and avoids some of the worst slurs and threats seen with MaRCo Infilling, although it still produces profanity and dehumanising language. DecompX Masking tends to lower toxicity compared to LLM Masking at the cost of more aggressive masking and slightly worse fluency. Overall, the reranker has the largest effect on safety, while masking and infilling control finer trade offs

between fluency, faithfulness, and residual toxicity.

Limitations

This study has several limitations. All experiments use a single English benchmark with a fixed reference style, so the results may not generalise to other domains or languages. The reranking signals are not learned jointly with the generators and rely on fixed scoring components. The toxicity classifier used for evaluation has its own biases and may misjudge indirect or context-dependent harm, and our qualitative analysis covers only a subset of examples without examining longer-term conversational effects. Finally, we operate under strict compute limits: we use only a moderate-sized LLM (Mistral 7B Instruct) for masking and infilling and cannot run larger models or much larger and more diverse datasets, which constrains both model capacity and evaluation breadth.

Future Work

Future work should learn rerankers that directly optimise a multi-objective detoxification criterion that combines toxicity, semantic preservation, and instruction fidelity. We also plan to explore richer masking strategies that blend attribution-based methods such as DecompX with contextual judgments from LLMs, and to systematically tune both the DecompX masking thresholds and the global reranking weights instead of fixing $w_T = 0.5$, $w_S = 0.3$, and $w_F = 0.2$.

On the generation side, we aim to replace Mistral 7B Instruct with stronger and more recent LLMs for both masking and infilling, such as ChatGPT 5.1, Gemini 3, and Opus 4.5. Finally, we plan to increase our compute capacity so that we can support larger LLMs and datasets and run broader evaluations across domains, languages, and human judgments of harm and usefulness.

Author Contributions

Benjamin He led the design of the T5-base models (T5-base, T5-base + DecompX Reranking, and T5-base + Global Reranking). He trained our T5-base model on ParaDetox and ran the experiments for the models and was responsible for creating the evaluation pipeline. He also primarily wrote the Introduction, Background, and Methods sections and edited the Results and Conclusion sections.

Kent Bourgoing designed the core modular pipeline based on the XDetox paper and implemented Jupyter notebooks to make experiments easy to run. He was responsible for the masking and infilling variants as well as for creating the new Global Reranking method. He also primarily wrote the Results and Conclusion sections and edited the Introduction, Background, and Methods sections.

Both authors discussed the ideas throughout the project, interpreted the findings together, and approved the final version of the paper.

References

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A Discourse-aware Transformer-based Style Transfer Model for Offensive Social Media Conversations. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*.

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 2023), Volume 2: Short Papers*.

Sara Bourbour Hosseinbeigi, Amin Saeidi Kelishami, Maryam Gheysari, and Fatemeh Rahimzadeh. 2025. MetaDetox at TextDetox CLEF 2025: Detoxification with Few-Chain Prompting. In *CLEF 2025 Working Notes*.

Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. XDetox: Text Detoxification with Token-Level Toxicity Explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*.

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 2022), Volume 1: Long Papers*.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining Transformers Decisions by Propagating Token Decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers*.

Jing Yu, Yibo Zhao, Jiapeng Zhu, Wenming Shao, Bo Pang, Zhao Zhang, and Xiang Li. 2025. Text Detoxification: Data Efficiency, Semantic Preservation and Model Generalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.

Appendix

Component	Hyperparameters / definition
Generation and masking	
T5 single-candidate	max_length=128, num_beams=5, no_repeat_ngram_size=2, greedy over beams.
T5 multi-candidate	max_length=128, do_sample=True, top_k=50, top_p=0.95, temperature=1.0, $C = 10$ candidates per input.
MaRCo (ParaDetox runs)	top_k_gen=50, top_p=0.95, filter_p=1.0, rep_penalty=1.0, $\alpha_a = 1.5$, $\alpha_e = 4.75$, temperature=2.5, max_length=96.
LLM infilling (Mistral-7B)	num_candidates=10, llm_temperature=0.7, llm_top_p=0.95, llm_max_new_tokens=64, sampling enabled.
DecompX masking / reranking	decompx_threshold=0.20, decompx_batch_size_mask=16 (where used), DecompX run on RoBERTa toxicity classifier.
Scoring (global reranker)	
Toxicity	Predicted probability of the "toxic" class from the XLM-R large toxicity classifier, applied to each candidate sentence (lower is better).
Semantic similarity	Cosine similarity between LaBSE sentence embeddings of the model output and the human reference, linearly rescaled to $[0, 1]$ (higher is better).
Fluency	GPT-2 perplexity p on the candidate sentence, mapped to a normalized fluency score $S_{\text{flu}} \in [0, 1]$ using $S_{\text{flu}} = \text{clip}((p_{\max} - p)/(p_{\max} - p_{\min}))$ with $p_{\min} = 5$ and $p_{\max} = 300$ (higher is better).
Global score	Global reranker score $\text{Score} = w_T(1 - p_{\text{tox}}) + w_S S_{\text{sim}} + w_F S_{\text{flu}}$ with weights $(w_T, w_S, w_F) = (0.5, 0.3, 0.2)$. The candidate with the highest score is selected.

Table 2: Generation, masking, and scoring configuration used in our experiments.

Prompt type	Prompt text
System prompt	You are a helpful assistant trained to make toxic or offensive sentences safer by MASKING toxic words with the special token <mask>, while keeping the original sentence structure and wording as much as possible. You must behave like a MASKER, not a full rewriter. Your task: 1) Identify toxic, offensive, or profane words or short phrases. 2) For each toxic span, replace the entire span with a single <mask> token. 3) There may be multiple toxic spans in one sentence, so multiple <mask> tokens are allowed. 4) If several neighboring words are toxic, you must still use only a single <mask> token in that place. In other words, if you would place "<mask> <mask>" or a longer sequence of <mask> tokens, collapse them into a single <mask> so that there are never multiple <mask> tokens in a row. 5) Do NOT rewrite, paraphrase, or summarize the sentence. 6) Do NOT add, remove, or reorder non-toxic words or punctuation. 7) Keep punctuation and spacing as close to the original as possible. 8) If there is no toxic content, return the sentence unchanged. Output rules (format is very strict): ONLY return the final masked sentence inside ONE pair of square brackets, like: [This is a <mask> example.]. Do NOT print anything before or after the brackets. Do NOT add explanations, comments, or extra lines. Do NOT include any language tags or metadata. Do NOT include additional '[' or ']' characters inside the sentence.
Few-shot prompt	Toxic Sentence: You're such a stupid idiot, nobody wants to hear your crap. Step 1 - Identify toxic words: "stupid idiot", "crap". Step 2 - Mask toxic words (do NOT rewrite the rest): You're such a <mask>, nobody wants to hear your <mask>. Final Output: [You're such a <mask>, nobody wants to hear your <mask>.]

Table 3: LLM masking prompts used for Mistral-7B-Instruct.

Prompt type	Prompt text
System prompt	You are a helpful assistant trained to make toxic or offensive sentences more polite and respectful by INFILLING the special token <mask>. You are NOT a free rewriter. You must keep all non-masked text as close as possible to the given masked sentence. You are given two inputs: 1) Toxic Sentence: the original toxic sentence. 2) Masked Sentence: the same sentence, where toxic spans are replaced with <mask>. Your task: 1) For each <mask> token in the Masked Sentence, replace it with a short, non-toxic word or phrase that fits the context and preserves the meaning of the Toxic Sentence. 2) Do NOT modify any other words or punctuation outside the <mask> spans, unless a very small change is needed to fix grammar or agreement. 3) Preserve the original meaning and intent as much as possible, but make the sentence safe and respectful. 4) Keep the language the same as the original (do NOT translate). Output rules (VERY STRICT): ONLY return the final detoxified sentence with all <mask> tokens filled. Wrap the final sentence in exactly ONE pair of square brackets, e.g.: [Detoxified sentence here.]. Do NOT include the Toxic Sentence or Masked Sentence in your output. Do NOT add explanations, comments, or extra lines. Do NOT include any other '[' or ']' characters.
Few-shot prompt	Toxic Sentence: You're such a stupid idiot, nobody wants to hear your crap. Masked Sentence: You're such a <mask>, nobody wants to hear your <mask>. Step 1 - Decide safe replacements for each <mask>: "rude person", "opinion". Step 2 - Infill the masked sentence, keeping all other words the same: You're such a rude person, nobody wants to hear your opinion. Final Output: [You're such a rude person, nobody wants to hear your opinion.]

Table 4: LLM infilling prompts used for Mistral-7B-Instruct.

Original	DecompX-masked	LLM-masked
. or the loud ass one - thousand ton beast roaring towards you howling its horn .	. or the loud <mask> one - thousand ton beast roaring towards you howling its horn .	. or the loud <mask> one - thousand ton beast roaring towards you howling its horn .
mandated and " right fucking now " would be good .	mandated and " right <mask> now " would be good .	mandated and "<mask> right <mask> now <mask> " would be good .
6. are you fucking insane ?	6. are you <mask> <mask> ?	6. are you <mask> ?
fuck ever doing that again .	<mask> ever doing that again .	fuck ever doing that again .
minorities bitch and moan when the police uphold the law .	minorities <mask> and moan when the police uphold the law .	minorities bitch and moan when the police uphold the law .

Table 5: Example original inputs and masked versions from DecompX and the LLM-based masker. Rows illustrate both successful and unsuccessful masking behaviour; some cells contain toxic language.

Reference / model	Example outputs
Reference (human paraphrases)	1) You are so annoying sometimes, just be quiet for once, I am tired of your constant complaining. 2) Well, he fatally shot himself. 3) How can you be so inconsiderate, leaving your mess for others to clean up?
T5-base	1) You are so annoying sometimes, just shut up for once, I am tired of your constant whining. 2) *Well, he shot himself. 3) Are you really that selfish, dumping your mess on others to clean up?
T5-base + DecompX Reranking	1) You are so annoying sometimes, just shut up for once, I am tired of your constant whining. 2) *Well, he shot himself. 3) Are you really that selfish, dumping your mess on others to clean up?
T5-base + Global Reranking	1) You can be frustrating at times, please be quiet for a moment, I am exhausted by your continuous complaints. 2) Unfortunately, he tragically harmed himself. 3) It is surprising how inconsiderate it can seem to leave a mess for others to handle.
DecompX Masking + MaRCo Infilling + DecompX Reranking	1) You are so annoying sometimes, I am tired of your constant whining. 2) *Well, he shot himself. 3) Are you really that selfish, dumping your mess on others to clean up?
DecompX Masking + MaRCo Infilling + Global Reranking	1) You are frustrating at times, please be quiet for a moment, I am exhausted by your continuous complaints. 2) Unfortunately, he tragically harmed himself. 3) It is surprising how inconsiderate it can seem to leave a mess for others to handle.
DecompX Masking + LLM Infilling + DecompX Reranking	1) You are so frustrating sometimes, please keep quiet for a moment, I am exhausted from your constant complaints. 2) Sadly, he ended his own life. 3) It is surprising how inconsiderate it can be to leave your clutter for others to tidy up.
DecompX Masking + LLM Infilling + Global Reranking	1) You can be frustrating at times, please keep quiet for a moment, I am exhausted from your constant complaints. 2) Sadly, he ended his own life. 3) It is surprising how inconsiderate it can be to leave your clutter for others to tidy up.
LLM Masking + MaRCo Infilling + DecompX Reranking	1) You are so annoying sometimes, just be quiet for once, I am tired of your constant complaining. 2) *Well, he shot himself. 3) You are so selfish, leaving your mess for everyone else to clean up.
LLM Masking + MaRCo Infilling + Global Reranking	1) You can be frustrating at times, please be quiet for a moment, I am exhausted by your continuous complaints. 2) Unfortunately, he tragically harmed himself. 3) It is surprising how inconsiderate it can seem to leave a mess for others to handle.
LLM Masking + LLM Infilling + DecompX Reranking	1) You can be frustrating at times, please keep quiet for a moment, I am exhausted from your constant complaints. 2) Sadly, he ended his own life. 3) It is surprising how inconsiderate it can be to leave your clutter for others to tidy up.
LLM Masking + LLM Infilling + Global Reranking	1) You can be frustrating at times, please keep quiet for a moment, I am exhausted from your constant complaints. 2) Sadly, he ended his own life. 3) It is surprising how inconsiderate it can be to leave your clutter for others to tidy up.

Table 6: Three reference sentences and corresponding model outputs for all eleven pipelines.

Input / model	Example outputs (3 sentences)
Original toxic sentences	1) You sicken me with that disgusting abortion, how can you even consider it? 2) Ugh, here we go again supporting people who hate people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is a piece of shit.
T5-base	1) Abortion is not a fundamental right, nothing more. 2) Ah, another supporter for those who hate people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is not good.
T5-base + DecompX Reranking	1) Abortion is not a fundamental right, it is merely wrong. Nothing more. 2) Ah, another supporter for those who hate people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is useless.
T5-base + Global Reranking	1) Abortion is not a fundamental right; it is purely nonsense. 2) Another supporter for those who hate people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is not that good.
DecompX Masking + MaRCo Infilling + DecompX Reranking	1) Comparing that abortion to a medical procedure is disgusting, how can you even say that? 2) Ugh, here we go again supporting those who oppose people of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is a waste of space.
DecompX Masking + MaRCo Infilling + Global Reranking	1) Comparing that abortion to a medical procedure shows a serious lack of understanding, how can you even say that? 2) Ugh, here we go again supporting people who oppose individuals of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is not a suitable candidate.
DecompX Masking + LLM Infilling + DecompX Reranking	1) You sicken me with that decision, how can you even consider it? 2) Ugh, here we go again supporting those who discriminate against people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still feel this guy is not the best choice.
DecompX Masking + LLM Infilling + Global Reranking	1) You sicken me with that decision, how can you even consider it? 2) Ugh, here we go again supporting those who discriminate against people of color and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still feel this guy is not the best choice.
LLM Masking + MaRCo Infilling + DecompX Reranking	1) Abortion is a horrible choice, how can you even think about it? 2) Ugh, here we go again supporting those who hate people of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is not worth supporting.
LLM Masking + MaRCo Infilling + Global Reranking	1) Abortion is a serious decision, how can you even consider it? 2) Ugh, here we go again supporting those who oppose people of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still say this guy is not a great candidate.
LLM Masking + LLM Infilling + DecompX Reranking	1) Abortion is a difficult decision, how can you even consider it? 2) Ugh, here we go again supporting those who discriminate against people of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still feel this guy is not a suitable choice.
LLM Masking + LLM Infilling + Global Reranking	1) Abortion is a deeply personal decision, how can you even consider it? 2) Ugh, here we go again supporting those who discriminate against people of different races and those of the Jewish faith. 3) As someone from LA and an oilfield worker, I still feel this guy is not the best choice.

Table 7: Three original toxic inputs and corresponding model outputs for all eleven pipelines.