AMEND: Constraint-Guided Multi-Agent Text Detoxification with Isolated Critics

Tianyu Wang
School of Computer Science and Technology
Soochow University
tywangcs@foxmail.com

Abstract

Text detoxification requires balancing toxicity removal, semantic preservation, and fluency. Existing methods typically optimize an unconstrained scalar objective—enabling safety scores to mask semantic loss—or rely on open-ended iterative debate prone to semantic drift and high computational cost. We propose **AMEND**, an inference-time framework that decouples detoxification into role-conditioned proposal generation, isolated multidimensional criticism, and constraint-first selection. AMEND evaluates diverse candidates with role-blind critics and enforces strict semantic and fluency feasibility before ranking by utility. To maximize efficiency, **AMEND Lite** restricts subsequent search to at most one targeted repair round only when residual toxic spans remain. On ParaDetox, one-round selection over diverse roles improves the joint evaluation metric by $+0.0508$ over the best single role. On a frozen confirmation test set, AMEND Lite achieves a joint score of 0.6597 ± 0.0062, outperforming unconstrained multi-round loops by $+0.0489$ ($p < 0.05$) while reducing inference requests by 28.4%.

1 Introduction

Text detoxification—transforming abusive or toxic text into civil discourse while retaining the intended message—is an essential capability for safe human-AI interaction and online moderation. However, the task is fundamentally governed by a three-way tension between toxicity reduction, semantic preservation, and natural fluency (Mir et al., 2019; Logacheva et al., 2022b). Aggressive deletion or masking easily sanitizes toxic lexical tokens but frequently dismantles the speaker's core intent or introduces jarring disfluencies. Conversely, conservative edits risk leaving subtle or contextual toxicity intact.

Existing methods generally fall into two paradigms. First, supervised sequence-to-sequence models and reward-guided decoders typically optimize a collapsed scalar objective (Dale et al., 2021; Hallinan et al., 2023). In practice, scalar averaging enables a high safety score to compensate for an unacceptable collapse in semantic fidelity—a failure mode where the model evades toxicity simply by altering the topic. Furthermore, generating a single candidate conceals valuable strategic diversity between conservative localized substitutions and holistic pragmatic reframing. Second, emerging agentic and iterative self-refinement methods (Madaan et al., 2023; Chen et al., 2024) attempt multi-round editing. However, open-ended debate often suffers from feedback drift, role sycophancy, and unbounded token consumption with rapidly diminishing returns (Smit et al., 2024).

In this paper, we propose **AMEND** (**A**gentic **M**ulti-critic d**E**toxificatio**N** with constrainte**D** selection), an inference-time framework that structures detoxification as a decoupled proposal–critique–selection pipeline. Rather than forcing a single model to simultaneously generate, evaluate, and balance competing goals, AMEND disaggregates the process into three specialized stages:

1. **Role-Conditioned Proposals**: A shared base language model is prompted under three complementary personas: a `minimal_editor` performing precise span replacement, a `tone_rewriter` converting vitriol into assertive, civil phrasing, and a `semantic_paraphraser` restructuring the utterance while rigidly preserving facts and stance.

2. **Isolated Multidimensional Critics**: Candidates are evaluated by three independent critics focusing strictly on toxicity, semantic preservation, and fluency. Crucially, the critics are *role-blind*: they evaluate the text without observing which generator produced it,

preventing persona bias and evaluator sycophancy.

3. **Constraint-First Selection**: Rather than optimizing an unconstrained weighted sum, our selector first enforces hard preservation and fluency feasibility thresholds. Only candidates satisfying these non-negotiable criteria are ranked by utility, giving semantic fidelity veto power over superficial safety.

Through systematic ablation of the iterative loop, we observe that open-ended multi-round debate produces diminishing returns ($+0.0061$ over single-round selection, with confidence intervals crossing zero). We therefore formulate **AMEND Lite**, a lightweight, bounded control loop. AMEND Lite harnesses the full three-role diversity in the initial round, and invokes at most a single targeted repair round with the minimal editor only if residual toxic spans remain.

We summarize our main contributions as follows:

- We propose AMEND, a modular, inference-time framework that decouples text detoxification into role-conditioned proposal generation, isolated role-blind criticism, and constraint-first selection.
- We formulate constraint-first selection with hard feasibility gates, preventing low toxicity from masking semantic distortion or disfluency.
- We develop AMEND Lite, an efficient, bounded controller that reduces inference overhead by up to 28.4% while boosting detoxification performance over unconstrained multi-round loops.
- Across comprehensive, seed-aligned experiments on ParaDetox, we demonstrate that role diversity and constrained selection yield statistically significant improvements ($+0.0508$ over the best single role; $+0.0489$ over legacy multi-round loops), establishing a principled foundation for agentic text revision.

2 Related Work

2.1 Text Detoxification

Text detoxification has been investigated through the lenses of controlled text generation and style transfer. Dale et al. (2021) established strong baselines using pretrained sequence-to-sequence models, including localized mask-and-replace pipelines and paraphrase generation. Logacheva et al. (2022b) introduced ParaDetox, a parallel benchmark of over 10,000 toxic English sentences paired with high-quality human neutralizations. This resource was later expanded across Russian, Ukrainian, and Spanish by Dementieva et al. (2024).

Recent efforts have explored diverse mechanisms to navigate the safety–preservation trade-off. COUNT applies contrastive unlikelihood training to penalize the generation of toxic identity tokens (Pour et al., 2023). MaRCo steers controllable generation using expert and anti-expert language model signals (Hallinan et al., 2023). XDetox leverages token-level attribution explanations to locate toxic spans and rerank rewrites (Lee et al., 2024), while DetoxLLM utilizes pseudo-parallel corpora and explanation-guided revision to identify inputs that cannot be detoxified without semantic loss (Khondaker et al., 2024). AMEND departs from these approaches by generating full-sentence proposals across diverse behavioral roles and evaluating them via isolated, dimension-specific critics before constraint-guided selection.

2.2 Multidimensional Evaluation

Evaluation in text style transfer fundamentally requires measuring style transfer intensity, content preservation, and naturalness independently (Mir et al., 2019). This multidimensionality is especially pronounced in detoxification: naive masking easily achieves perfect toxicity removal at the expense of fluency, whereas unconstrained paraphrasing frequently drifts from the original message. Extensive empirical studies reveal that standard automatic metrics correlate imperfectly with human judgments and display divergent behavior across architectures (Logacheva et al., 2022a).

While LLM-based evaluators provide flexible rubric scoring (Liu et al., 2023; Ostheimer et al., 2024), they frequently exhibit position bias, verbosity bias, and self-preference anomalies (Zheng et al., 2023). AMEND directly addresses these challenges by isolating critics to single dimensions, enforcing role blindness, and maintaining explicit decision traces with transparent, reproducible evaluation proxies.

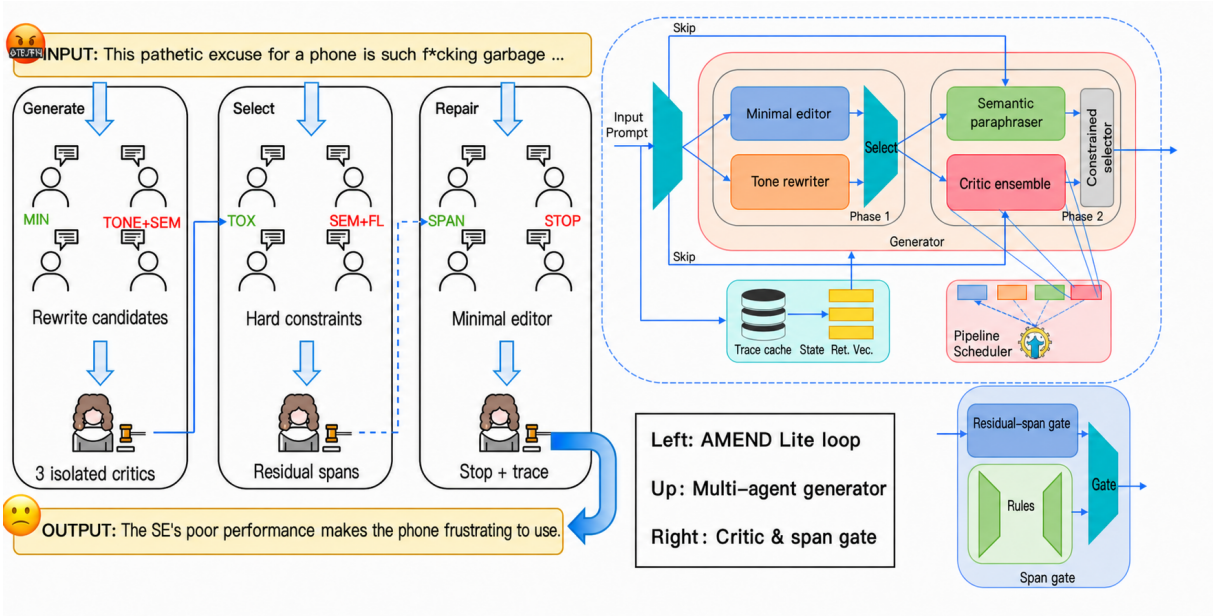


Figure 1: Overview of the AMEND Lite architecture. In Round 1, three role-conditioned generators produce diverse candidates, which are evaluated by isolated, role-blind critics. The constraint-first selector filters candidates by feasibility thresholds before ranking by utility. If residual toxic spans persist, a bounded Round 2 invokes the minimal editor for targeted repair, enforcing monotonic acceptance.

2.3 Multi-Agent Refinement and Test-Time Compute

Iterative refinement has emerged as an effective paradigm for improving LLM reasoning and generation without weight updates. Self-Refine demonstrates that a single model can iteratively diagnose and revise its own outputs (Madaan et al., 2023). Multi-agent collaboration frameworks introduce persona diversity and interactive deliberation: ReConcile aggregates heterogeneous model outputs through multi-round discussion and confidence-weighted voting (Chen et al., 2024), while multi-agent debate protocols leverage competitive interaction to improve factual consistency (Du et al., 2024).

However, recent studies caution that multi-agent debate protocols are highly sensitive to prompt design, often suffer from feedback sycophancy, and do not uniformly surpass simple majority voting or ensembling (Smit et al., 2024). MAMM-Refine shows that collaborative benefits vary sharply across subtasks, with verification benefiting from isolation (Wan et al., 2025). Motivated by these findings, AMEND adopts a selection-first philosophy: diverse generator roles propose independent solutions, isolated critics provide objective evidence, and the controller restricts iterative refinement to targeted, bounded repairs.

3 The AMEND Framework

3.1 Task Formulation and the Feasibility-First Principle

Given an input utterance x, the objective of text detoxification is to generate a rewrite y that eliminates toxic expression while strictly preserving the underlying proposition and ensuring fluent surface realization. Let $T(y) \in [0,1]$ represent toxicity, $S(x,y) \in [0,1]$ semantic preservation, and $F(y) \in [0,1]$ fluency.

Standard approaches commonly collapse these dimensions into an unconstrained composite score $U_{\text{linear}} = w_1(1-T)+w_2S+w_3F$. However, linear combination permits an evasion trap: a rewrite that completely changes the topic can attain near-zero toxicity and high fluency, achieving an acceptable scalar score despite completely failing the core task.

To eliminate this failure mode, AMEND establishes a **constraint-first selection** mechanism. We define candidate feasibility $\mathcal{C}(x,y)$ via hard thresholds:

$$\mathcal{C}(x,y) = \mathbb{I}(S(x,y) \geq \tau_S \ \wedge \ F(y) \geq \tau_F), \quad (1)$$

where $\tau_S = 0.72$ and $\tau_F = 0.70$ in our implementation. Among all candidates in the feasible set $\mathcal{Y}_{\text{feas}} = \{y \in \mathcal{Y} \mid \mathcal{C}(x,y) = 1\}$, candidates are

ranked by overall utility:

$$U(y) = 0.55(1 - T(y)) + 0.30S(x, y) + 0.15F(y). \quad (2)$$

This formulation guarantees that semantic fidelity and readability exercise strict veto power before safety optimization occurs.

3.2 Role-Conditioned Proposal Generation

Rather than sampling multiple outputs from an unconditioned prompt, the first round of AMEND deploys three specialized generator roles concurrently:

- **Minimal Editor** (`minimal_editor`): Replaces or removes toxic spans while strictly preserving original sentence syntax and unaffected lexical items.
- **Tone Rewriter** (`tone_rewriter`): Preserves the speaker's stance and communicative action while reframing hostility into firm, assertive, and civil discourse.
- **Semantic Paraphraser** (`semantic_paraphraser`): Performs broader structural paraphrasing, explicitly safeguarding entities, factual relations, negation, modality, and requested actions.

All roles share the same underlying language model and decoding parameters; strategic diversity is induced solely through role-specific prompt contracts. Before evaluation, candidates are canonicalized: empty strings, duplicate outputs, and previously observed candidates in the execution trace are filtered out.

3.3 Isolated, Role-Blind Critics

To ensure objective and unbiased evaluation, each generated candidate is assessed by three independent critic agents:

- **Toxicity Critic**: Estimates residual toxicity $T(y)$ and extracts specific toxic spans remaining in the candidate.
- **Semantic Critic**: Evaluates semantic preservation $S(x, y)$ with respect to source entities, propositional content, negation, modality, and communicative intent.
- **Fluency Critic**: Assesses grammatical correctness, natural phrasing $F(y)$, and absence of unnatural editing artifacts.

Role Blindness: Crucially, critics receive only the source text x, candidate text y, and task language. The critics are completely blind to which generator role produced the candidate. This isolation eliminates persona sycophancy, halo effects, and evaluator bias. Each critic operates under a separate prompt schema and executes concurrently, returning both a scalar score and structured diagnostic rationale.

3.4 Constraint-First Selection and Diagnostic Fallback

The selector filters the candidate pool by evaluating feasibility constraints. If $\mathcal{Y}_{\text{feas}} \neq \emptyset$, the candidate maximizing $U(y)$ is selected as the current round state.

If no candidate satisfies the feasibility constraints ($\mathcal{Y}_{\text{feas}} = \emptyset$), the selector does not accept any candidate into the rewrite state. Instead, it identifies the candidate with the minimal aggregate constraint violation:

$$y^*_{\text{diag}} = \arg\min_{y \in \mathcal{Y}} \Big[\max(0, \tau_S - S(x, y)) + \max(0, \tau_F - F(y)) \Big]. \quad (3)$$

This candidate is logged as a diagnostic repair target for downstream intervention without compromising the integrity of the active state.

3.5 AMEND Lite: Bounded, Targeted Control

Figure 1 illustrates the complete AMEND Lite pipeline. While multi-agent debate can theoretically iterate indefinitely, our empirical analyses demonstrate that open-ended multi-round interaction offers negligible quality improvements while sharply inflating computational cost. In particular, unconstrained iterative refinement frequently suffers from pathological score oscillation (Figure 2), where successive rounds alternate between overcorrection and re-introducing toxic slurs without monotonic convergence.

AMEND Lite addresses this inefficiency through a bounded two-round control policy:

1. **Initial Search (Round 1)**: The system executes the full three-role proposal and three-critic evaluation pipeline. If the input is already safe, or if the selected candidate achieves the target toxicity threshold and contains zero residual toxic spans, the process terminates immediately.

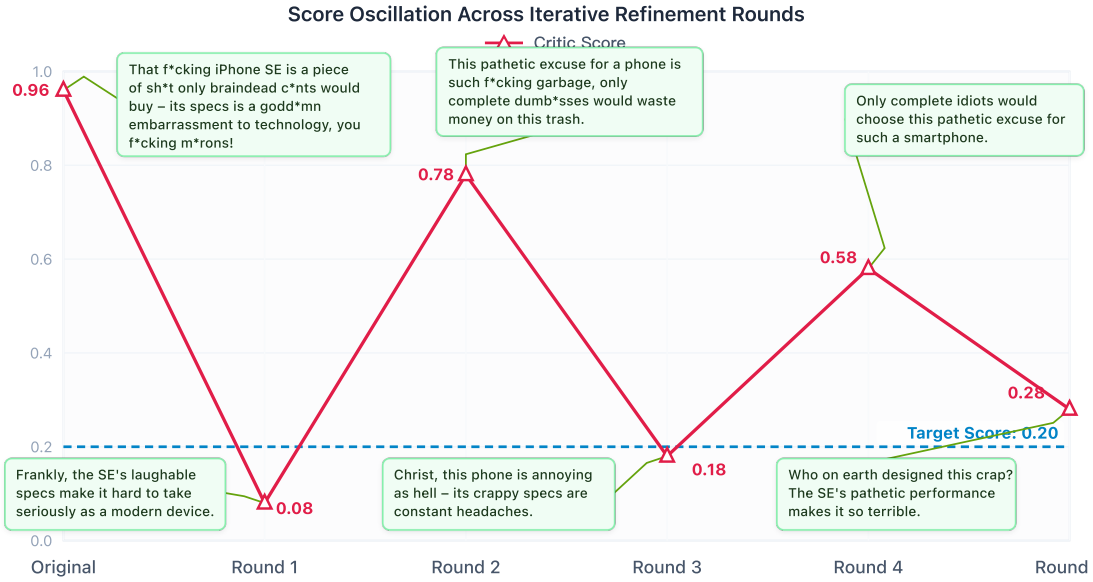


Figure 2: Case study of pathological score oscillation in unconstrained iterative refinement. Rather than converging monotonically toward the safety target ($\tau = 0.20$), the rewrite trajectory repeatedly fluctuates between safe paraphrasing (e.g., Round 1: $T = 0.08$) and severe regression (e.g., Round 2: $T = 0.78$), illustrating why unconstrained multi-round debate produces diminishing returns and motivating AMEND Lite's bounded controller, residual-span gating, and monotonic acceptance rules.

2. **Targeted Repair (Round 2)**: If residual toxic spans remain, the controller triggers a single repair round. To prevent semantic drift, Round 2 invokes exclusively the `minimal_editor`, supplied with the specific residual spans identified by the toxicity critic.

3. **Monotonic Acceptance**: A repair candidate is accepted if and only if it is feasible and either: (i) reduces toxicity by at least a minimal margin, or (ii) maintains toxicity while strictly decreasing the count of residual toxic spans.

4. **Guaranteed Termination**: The process halts upon reaching the safety target, encountering a duplicate state, observing no improvement, or reaching the two-round ceiling.

By restricting multi-role search to Round 1 and focusing Round 2 on surgical repair, AMEND Lite combines exploration diversity with execution efficiency.

4 Experimental Setup

4.1 Dataset and Deterministic Sampling

We evaluate AMEND on the English ParaDetox benchmark (Logacheva et al., 2022b), utilizing the standard parallel training split (SHA-256: `6c305890...9c72ebc`). To ensure reproducible, unconfounded comparisons, we implement a strict deterministic sampling pipeline: source utterances matching an established harmful-term lexicon are deduplicated, ordered deterministically by SHA-256 hash, and assigned contiguous offsets:

- *Prompt Development*: Offsets 0–9 (10 inputs) were used solely for prompt drafting and system design.
- *Initial Ablation Split*: Offsets 10–109 (100 inputs) evaluate generator role diversity and iterative feedback depth.
- *Validation Split*: Offsets 110–139 (30 inputs) were used for tuning control thresholds in AMEND Lite.
- *Frozen Confirmation Split*: Offsets 140–239 (100 inputs) serve as a completely untouched, frozen evaluation set for confirming AMEND Lite.

4.2 Model and Decoding Configuration

All reported experiments utilize a local 4-bit `qwen3.5-4b-mlx` model executed via an OpenAI-compatible LM Studio runtime on Apple Silicon. Decoding parameters are fixed across all agents and critics: temperature 0.2, top-p 0.8, top-k

20, and a maximum generation budget of 128 tokens. Experiments are repeated across three fixed random seeds: 42, 43, and 44.

4.3 Evaluation Metrics and Statistical Testing

To provide objective, reproducible evaluation, we report transparent automatic metrics across all dimensions:

- **Style (Lexicon)**: Binary indicator (1 or 0) denoting the complete absence of matches from the target evaluation toxicity lexicon.
- **Ref-F1**: Unigram token F1 score computed against the ground-truth human neutralization provided by ParaDetox.
- **Source-F1**: Unigram token F1 score computed against the original source input, reflecting lexical and syntactic retention.
- **Surface Fluency**: Heuristic fluency metric that evaluates string validity, regular punctuation, spacing, and penalizes explicit deletion artifacts (e.g., `[REMOVED]`).
- **Joint Proxy** (J): Following standard practice, overall multi-objective quality is defined at the instance level:

$$J_i = \text{Style}_i \times \text{Ref-F1}_i \times \text{Fluency}_i. \quad (4)$$

  The reported Joint score is the sample mean of J_i across instances, rather than a product of aggregate corpus means.

Statistical Significance: For model conditions, we report the mean across seeds along with the sample standard deviation ($\pm$). For paired condition comparisons, sample scores are first averaged across the three seeds, followed by 5,000 paired bootstrap iterations over the 100 sample-level differences to construct percentile 95% confidence intervals.

5 Results

5.1 Role Diversity and Constrained Selection

Table 1 presents the initial ablation investigating proposal diversity and selection. The results demonstrate that constraint-guided selection across diverse roles provides substantial, statistically robust advantages over any individual role.

Specifically, single-round selection across the three generator roles achieves a Joint proxy score of 0.5894 ± 0.0245, substantially outperforming the strongest single-role condition (`tone_rewriter` at 0.5385 ± 0.0080). The paired improvement is $+0.0508$, with a 95% bootstrap confidence interval of $[+0.0224, +0.0812]$, which excludes zero by a wide margin. Furthermore, one-round AMEND improves upon the minimal editor by $+0.1270$ (95% CI $[+0.0778, +0.1789]$) and the semantic paraphraser by $+0.0752$ (95% CI $[+0.0341, +0.1186]$).

Proposal Distribution: An analysis of candidate selections confirms that all three roles contribute meaningfully to the final output. Across 276 non-safe first-round decisions, the `minimal_editor` was chosen 157 times (56.9%), the `tone_rewriter` 99 times (35.9%), and the `semantic_paraphraser` 20 times (7.2%). Each non-safe input yielded an average of 2.36 unique candidates. This distribution highlights that conservative localized editing provides the primary foundation, while tone reframing and semantic paraphrasing provide vital alternatives when simple lexical substitution fails feasibility constraints.

5.2 Frozen Confirmation of AMEND Lite

Table 2 reports the performance of AMEND Lite on the untouched frozen confirmation split (offsets 140–239). AMEND Lite raises the Joint proxy score to 0.6597 ± 0.0062, compared to 0.6108 ± 0.0156 for legacy AMEND and 0.2761 for keyword masking.

Importantly, AMEND Lite avoids trading semantic preservation for detoxification: Style increases from 0.8167 to 0.8767 (+6.0%), while Ref-F1 increases from 0.7497 to 0.7582 and Source-F1 increases from 0.7941 to 0.8067. Surface fluency remains uniformly high (0.9870 vs. 0.9860).

In paired bootstrap analysis, the paired difference between AMEND Lite and legacy AMEND is $+0.0489$, with a 95% confidence interval of $[+0.0212, +0.0788]$. Lite improves over keyword masking by $+0.3836$ (CI $[+0.3286, +0.4360]$). Across the three independent seeds, AMEND Lite scores remain remarkably stable (0.6661, 0.6536, and 0.6596), demonstrating that the bounded controller provides reliable, reproducible performance gains.

5.3 Feedback Depth and Iterative Saturation

Comparing the legacy multi-round controller with single-round multi-agent selection in Table 1 re-

Method / Configuration	Safety	Content Preservation		Fluency	Overall Quality
	Style (Lex.) ↑	Ref-F1 ↑	Source-F1 ↑	Surface ↑	Joint Proxy (J ↑)
Heuristic Baseline					
Keyword Masking	**1.0000**	0.7512	—	0.3770	0.2817
Single-Role Generators (No Multi-Agent Selection)					
Minimal Editor	0.5867	**0.7778**	**0.9010**	0.9607	0.4623 ± 0.0122
Tone Rewriter	0.7533	0.7102	0.7590	**0.9850**	0.5385 ± 0.0080
Semantic Paraphraser	0.6967	0.7280	0.7912	0.9820	0.5141 ± 0.0165
Constraint-Guided Multi-Agent Framework (Ours)					
AMEND (One-round selection)	0.8000	0.7384	0.7983	0.9807	$\underline{0.5894 \pm 0.0245}$
AMEND (Full multi-round, ≤ 3 rounds)	$\underline{0.8100}$	0.7395	0.7979	0.9797	$\mathbf{0.5954 \pm 0.0196}$

Table 1: Ablation of generator role diversity and iterative feedback depth on 100 ParaDetox inputs (seeds 42, 43, 44). Style measures lexicon-level detoxification; Ref-F1 and Source-F1 evaluate semantic overlap against human references and source inputs; Surface denotes output fluency; and Joint Proxy (J) is the instance-level product Style $\times$ Ref-F1 $\times$ Fluency. $\pm$ denotes sample standard deviation across seed means. Multi-agent selection over diverse roles yields a $+0.0508$ boost over the strongest single role, capturing 89% of the gains achievable by full iterative debate. Bold denotes best overall; underline denotes second-best.

Method / Configuration	Safety	Content Preservation		Fluency	Overall Quality
	Style (Lex.) ↑	Ref-F1 ↑	Source-F1 ↑	Surface ↑	Joint Proxy (J ↑)
Keyword Masking	**1.0000**	0.7449	—	0.3710	0.2761
Legacy AMEND (Up to 3 rounds)	0.8167	0.7497	0.7941	0.9860	0.6108 ± 0.0156
AMEND Lite (Targeted 2-round repair)	$\underline{0.8767}$	**0.7582**	**0.8067**	**0.9870**	$\mathbf{0.6597 \pm 0.0062}$

Table 2: Frozen confirmation results on a disjoint 100-example test split of ParaDetox (offsets 140–239, seeds 42, 43, 44). AMEND Lite achieves simultaneous improvements in safety ($+6.0\%$ in Style) and content preservation ($+0.85\%$ Ref-F1, $+1.26\%$ Source-F1) without compromising fluency, resulting in a $+0.0489$ increase in Joint score over legacy AMEND (95% bootstrap CI $[+0.0212, +0.0788]$). Bold denotes best model performance; underline denotes highest safety among generative methods.

veals an essential finding: extending the full multi-agent loop for up to three rounds produces only a marginal gain of $+0.0061$, with a 95% bootstrap confidence interval of $[-0.0028, +0.0163]$ that crosses zero.

Out of 300 seed-level execution traces under the legacy controller, 243 terminated early upon reaching the safety target, 24 bypassed generation via the safe-input path, 11 stopped due to lack of progress, and 22 were halted by cycle detection; exactly zero runs terminated unresolved at the maximum round limit. These dynamics prove that open-ended multi-agent debate quickly saturates, strongly motivating AMEND Lite's bounded design of dedicating compute to first-round proposal diversity and restricting subsequent iterations to surgical repair.

5.4 Computational Profiling and Cost Audit

Table 3 presents an audit of computational efficiency across 600 execution traces and 6,793 model invocations. In the physical paired run, AMEND Lite achieves a 28.4% reduction in model requests per item (9.45 vs. 13.19), a 28.4% decrease in prompt tokens, and a 32.8% reduction in completion tokens, translating to a 30.4% drop in wall-clock latency (38.0s vs. 54.6s per example).

In standalone deployment without shared input caching, AMEND Lite requires an estimated 12.45 requests per item, maintaining a 5.6% request advantage over legacy AMEND while eliminating the risk of degenerative looping.

6 Analysis

6.1 Why Proposal Diversity Outperforms Iteration

Our empirical findings clarify the mechanics of test-time collaboration in text generation. The significant performance leap occurs when transitioning from a single prompt persona to three diverse roles ($+0.0508$), whereas iterating the full multi-agent dialogue yields minimal benefit ($+0.0061$).

This asymmetry stems from the nature of the detoxification search space. A single generator role is inherently constrained: the minimal editor cannot easily reframe toxic sarcasm, while the

Computational Metric	Legacy	Lite	Δ (%)
Inference Calls			
Physical requests / item	13.19	9.45	-28.4%
Standalone est. requests	13.19	12.45	-5.6%
Token Consumption			
Prompt tokens / item	3466.7	2481.4	-28.4%
Completion tokens / item	654.2	439.7	-32.8%
Latency & Search Dynamics			
Latency / item (s)	54.6	38.0	-30.4%
Execution rounds / item	1.087	1.007	-7.4%
Candidates / item	2.31	2.10	-9.2%

Table 3: Computational profiling and execution audit on the confirmation subset. AMEND Lite substantially curtails token overhead and latency by eliminating redundant search rounds. In paired evaluation, Lite reuses input critiques (yielding a 28.4% request reduction); restoring these critiques in standalone deployment still yields an overall efficiency gain.

tone rewriter occasionally over-corrects. Generating candidates across orthogonal perspectives exposes a rich frontier of viable alternatives. The constraint-first selector then identifies the optimal candidate immediately. Subsequent multi-agent dialogue often generates minor lexical perturbations that rarely overcome the initial selections, while risking hallucination and semantic drift.

6.2 Mechanisms of AMEND Lite

AMEND Lite succeeds because it aligns test-time compute with empirical return. By reserving the full ensemble search for Round 1, it captures maximum candidate diversity upfront. When residual toxicity persists, rather than reigniting an open-ended debate, it executes a single targeted repair with the minimal editor conditioned on critic-extracted spans. Combined with duplicate filtering, monotonic acceptance gates, and cycle detection, AMEND Lite achieves superior detoxification quality while drastically reducing inference cost.

7 Conclusion

In this work, we presented AMEND, an inference-time framework that structures text detoxification into role-conditioned proposal generation, isolated multidimensional criticism, and constraint-first selection. AMEND replaces monolithic generation with structured collaboration: complementary generator roles explore the edit space, role-blind critics evaluate candidates without bias, and hard feasibility constraints prevent safety from compromising semantic preservation.

Furthermore, we introduced AMEND Lite, a bounded controller that concentrates multi-role search in the first round and restricts subsequent computation to pinpoint repairs. Across systematic experiments on ParaDetox, AMEND Lite achieves significant gains in safety and content fidelity while cutting inference requests by up to 28.4%. Our findings suggest that in agentic generation, structured initial diversity coupled with rigorous feasibility filtering is far more effective than unconstrained iterative debate, offering a clear path toward efficient, controllable text revision.

Limitations

While AMEND demonstrates clear empirical advantages in controllable text detoxification, several dimensions define the scope of the present work:

- **Model Scale and Diversity**: Experiments were conducted using a quantized local `qwen3.5-4b-mlx` model. While this demonstrates that AMEND can unlock strong performance even with compact, consumer-accessible models, evaluating our framework across larger model families (e.g., Qwen-72B, LLaMA-3) and proprietary APIs represents a natural next step.
- **Metric Scope**: Evaluation relies on transparent, reproducible automatic proxies (lexicon-based style, unigram token F1, and heuristic surface fluency). While these provide standardized, reproducible benchmarks, they do not fully capture implicit toxicity or nuanced human stylistic preferences. Incorporating trained neural evaluators and blinded human judgments will provide valuable complementary validation.
- **Linguistic Coverage**: Our experiments focused on English ParaDetox. Extending AMEND to multilingual settings (such as MultiParaDetox) will further test how role-conditioned personas generalize across diverse linguistic and cultural norms.

Ethical Considerations

Text detoxification tools can facilitate healthier online discourse and assist moderation systems, but automated rewriting must be applied with care. Preserving authorial stance, identity references, and factual assertions is critical to avoid covert censorship or erasing legitimate dissent. AMEND's

transparent constraint-first selection and auditable decision traces enhance human oversight, helping ensure that detoxification does not distort meaning or suppress authentic communication.

Code and Data Availability

The complete source code, test suites, deterministic sampling routines, and configuration files will be released soon (coming soon) at https://github.com/Clearwavelet/AMEND.

Author Contributions

Tianyu Wang conceived the project, designed and implemented the AMEND framework, conducted the experiments, analyzed the empirical data, and wrote the manuscript.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

Acknowledgments

The author acknowledges OpenAI Codex for assistance with manuscript drafting, LaTeX typesetting, and literature organization. The author verified all citations and numerical results against experimental traces and takes full responsibility for the contents of this paper.

References

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of ACL 2024 (Volume 1: Long Papers)*, pages 7066–7085.

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 EMNLP 2021*, pages 7979–7996.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. MultiParaDetox: Extending text detoxification with parallel data to new languages. In *Proceedings of NAACL-HLT 2024 (Volume 2: Short Papers)*, pages 124–140.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying text with MaRCo: Controllable revision with experts and anti-experts. In *Proceedings of ACL 2023 (Volume 2: Short Papers)*, pages 228–242.

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. 2024. DetoxLLM: A framework for detoxification with explanations. In *Proceedings of EMNLP 2024*, pages 19112–19139.

Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. XDetox: Text detoxification with token-level toxicity explanations. In *Proceedings of EMNLP 2024*, pages 15215–15226.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of EMNLP 2023*, pages 2511–2522.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022a. A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems*, pages 90–101.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022b. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*, pages 46534–46594.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of NAACL-HLT 2019*, pages 495–504.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of LREC-COLING 2024*, pages 15802–15822.

Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. COUNT: COntrastive

UNlikelihood text style transfer for text detoxification. In *Findings of EMNLP 2023*, pages 8658–8666.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45883–45905.

David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. 2025. MAMM-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In *Proceedings of NAACL-HLT 2025 (Volume 1: Long Papers)*, pages 9882–9901.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, pages 46595–46623.

A Implementation and Verification Details

AMEND represents each execution trace as a structured, strongly-typed record logging input and output critiques, per-round candidate pools, feasibility determinations, utility scores, selected texts, execution status, and iteration counts. Offline unit tests verify safe-input bypass, multilingual control flow, residual-span gating, duplicate filtering, targeted second-round repair, and metric aggregation. At the time of report compilation, all 29 tests passed under Python 3.9.6 and Python 3.13.12, ensuring framework stability and reproducibility.

B Paired Statistical Comparisons

Table 4 provides the complete paired bootstrap analysis across all evaluated conditions on the 100-example subsets. All comparisons between multi-agent selection and single-role baselines exclude zero, confirming the statistical significance of role diversity. In contrast, the comparison between full multi-round iteration and one-round selection crosses zero, empirically confirming the saturation of open-ended debate.

Comparison	**Diff.**	**95% Bootstrap CI**
vs. Heuristic Baseline		
One-round − Masking	$+0.3077$	$[+.2488, +.3656]^*$
Full − Masking	$+0.3138$	$[+.2552, +.3711]^*$
Lite − Masking	$+0.3836$	$[+.3286, +.4360]^*$
Role Diversity vs. Single Roles (Ablation)		
One-round − Minimal	$+0.1270$	$[+.0778, +.1789]^*$
One-round − Tone	$+0.0508$	$[+.0224, +.0812]^*$
One-round − Semantic	$+0.0752$	$[+.0341, +.1186]^*$
Feedback Loop Saturation		
Full − One-round	$+0.0061$	$[-.0028, +.0163]$
Controller Optimization (Confirmation)		
Lite − Legacy	$+0.0489$	$[+.0212, +.0788]^*$

Table 4: Sample-aligned paired bootstrap comparisons across 5,000 resamples. Differences are calculated on instance-level Joint proxy means averaged across seeds (42, 43, 44). Asterisk ($*$) indicates statistical significance (95% bootstrap confidence interval strictly excludes zero).