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 must remove harmful expression while preserving meaning and producing natural language. We present AMEND, an inference-time framework that separates this objective into role-conditioned proposal generation, isolated multidimensional criticism, constraint-first selection, and bounded repair. Three generator roles propose minimal edits, tone-aware rewrites, and semantic paraphrases. Independent toxicity, semantic-preservation, and fluency critics evaluate each candidate without observing its generator role. A constrained selector first enforces semantic and fluency thresholds and then ranks feasible candidates by a transparent utility. AMEND Lite retains the diverse first-round proposal set and invokes one targeted repair only when residual toxic spans remain. In a frozen evaluation on 100 ParaDetox inputs with three seeds and a local 4-bit Qwen3.5-4B-MLX model, AMEND Lite obtains a joint proxy score of 0.6597 ± 0.0062, compared with 0.6108 ± 0.0156 for the legacy loop. The paired difference is $+0.0489$, with a 95% bootstrap interval of $[+0.0212, +0.0788]$. A separate ablation shows that one-round selection over three role-conditioned candidates improves over the strongest single role by $+0.0508$ on the same joint proxy. Diverse proposals and constraint-guided selection account for the principal measured gains.

1 Introduction

Text detoxification rewrites harmful language into a safer form while retaining the source message. The task is intrinsically multi-objective: a rewrite can remove overtly toxic words yet distort the speaker's proposition, or preserve the proposition while leaving harmful spans intact. Naturalness adds a third requirement because deletion and masking can produce conspicuous editing artifacts. These interacting objectives motivate evaluation along style, content, and fluency dimensions (Mir et al., 2019; Logacheva et al., 2022b).

Existing detoxification systems span localized replacement, controlled generation, paraphrase-based transfer, explanation-guided rewriting, and candidate reranking (Dale et al., 2021; Hallinan et al., 2023; Lee et al., 2024; Khondaker et al., 2024). A recurring design choice is how to combine competing signals. A single scalar objective permits a strong score on one dimension to compensate for an unacceptable failure on another. A single generated rewrite also hides useful variation between conservative editing and broader paraphrasing.

AMEND addresses these issues as an inference-time proposal–critique–selection process. The same base language model is assigned three generation roles with distinct editing priorities. Each candidate is then scored by three separately prompted critics, each restricted to one dimension. The selector treats semantic preservation and fluency as feasibility constraints before ranking candidates for safety and overall quality. The resulting trace records the proposals, critic outputs, constraint decisions, accepted edits, and stopping reason.

We further introduce AMEND Lite, a bounded control loop derived from an ablation of the original multi-round implementation. The first round retains all three generator roles and all three critics. A second round is invoked only when residual spans require targeted repair, and only the minimal editor is used. Duplicate and previously visited states are filtered before scoring, while monotonic acceptance and cycle detection prevent unproductive iteration.

This report makes three contributions. First, it presents an implemented architecture for role-diverse detoxification with role-blind, dimension-specific critics. Second, it formulates candidate choice as constraint-first selection, preventing low toxicity from compensating for inadequate semantic preservation or fluency. Third, it evaluates the proposal set, feedback depth, and Lite controller

through paired experiments on deterministic ParaDetox subsets. The results identify diverse first-round candidates and constrained selection as the main source of improvement in the evaluated proxy setting.

2 Related Work

2.1 Text Detoxification

Text detoxification has been studied as controlled text generation and style transfer. Dale et al. (2021) introduced strong pretrained-model approaches including localized mask-and-replace and paraphrase-guided generation. ParaDetox later provided more than 10,000 toxic sentences paired with human neutralizations, establishing a parallel benchmark for English detoxification (Logacheva et al., 2022b). MultiParaDetox extended parallel collection and evaluation to Russian, Ukrainian, and Spanish (Dementieva et al., 2024).

Subsequent systems target the safety–preservation trade-off through different mechanisms. COUNT discourages copying toxic identity content with a contrastive unlikelihood objective (Pour et al., 2023). MaRCo combines expert and anti-expert signals during controllable revision (Hallinan et al., 2023). XDetox locates toxic spans, generates replacements, and reranks candidates using token-level explanations (Lee et al., 2024). DetoxLLM constructs pseudo-parallel data, produces explanations and rewrites, and detects inputs that cannot be detoxified without changing their substance (Khondaker et al., 2024). AMEND differs by generating full-sentence proposals under multiple roles and applying isolated critics before a constraint-first selector.

2.2 Multidimensional Evaluation

Style transfer evaluation commonly separates transfer intensity, content preservation, and naturalness (Mir et al., 2019). This separation is especially important for detoxification: masking can score well on lexical removal while producing unnatural output, and unconstrained paraphrasing can improve style at the cost of meaning. Automatic detoxification metrics correlate imperfectly with human judgments, and their behavior varies across models and evaluation dimensions (Logacheva et al., 2022a). LLM-based evaluators provide structured, rubric-conditioned scoring at scale (Liu et al., 2023; Ostheimer et al., 2024), while also exhibiting position, verbosity, and self-preference biases (Zheng et al., 2023). AMEND therefore keeps critic dimensions separate in the decision trace and reports transparent external proxies as distinct columns rather than presenting a critic score as ground truth.

2.3 Feedback and Multi-Agent Refinement

Self-Refine demonstrates test-time generation, feedback, and revision with a single model (Madaan et al., 2023). Multi-agent methods add proposal diversity or discussion. ReConcile aggregates heterogeneous model outputs through repeated discussion and confidence-weighted voting (Chen et al., 2024), while multi-agent debate uses iterative exchange to improve reasoning and factuality (Du et al., 2024). Controlled comparisons show that debate protocols are sensitive to interaction design and do not uniformly dominate simpler ensembles (Smit et al., 2024). MAMM-Refine separates detection, critique, reranking, and revision, finding that different subtasks benefit from different forms of collaboration (Wan et al., 2025). AMEND follows a selection-first design: roles independently propose rewrites, isolated critics produce dimension-specific evidence, and an editor is called again only for a bounded repair.

3 AMEND

3.1 Task Formulation

Given a source utterance x, AMEND produces a rewrite y that reduces toxicity while preserving the source semantics and maintaining fluent language. Let $T(y) \in [0,1]$ denote toxicity, $S(x,y) \in [0,1]$ semantic preservation, and $F(y) \in [0,1]$ fluency. Candidate feasibility is defined as

$$S(x,y) \geq \tau_S, \qquad F(y) \geq \tau_F, \tag{1}$$

with $\tau_S = 0.72$ and $\tau_F = 0.70$ in the implementation. Feasible candidates are ranked by

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

This decomposition gives preservation and readability veto power before utility ranking.

3.2 Role-Conditioned Proposals

The first round runs three generator roles concurrently. The `minimal_editor` replaces toxic spans while retaining syntax and lexical content. The `tone_rewriter` preserves stance and intended action while expressing the message in firm, respectful language. The

`semantic_paraphraser` permits broader restructuring but explicitly preserves entities, facts, negation, modality, stance, and requested actions. All roles use the same model and decoding configuration; diversity is induced solely by isolated instructions.

No role is privileged in subsequent evaluation. Candidates are normalized, and empty outputs, same-round duplicates, and texts already observed in the trace are removed before critic calls. This makes the candidate pool an auditable set of distinct editing strategies.

3.3 Isolated Critics

Three critics receive only the source text, candidate text, and language. The toxicity critic estimates remaining toxicity and returns residual toxic spans. The semantic critic evaluates whether entities, factual content, stance, negation, modality, and action requests are retained. The fluency critic evaluates grammaticality, naturalness, and visible editing artifacts. Each critic uses a separate prompt and structured output, and the calls execute concurrently. Concealing the generator role prevents role labels from entering the evaluation context.

Critic outputs serve two purposes. Their scalar values determine candidate feasibility and ranking, while their residual-span and rationale fields guide trace inspection and a possible repair round. The full per-round trace records critic scores, feasibility, utility, selected candidate, and termination state.

3.4 Constraint-First Selection

The selector first filters candidates by the semantic and fluency thresholds. It then maximizes $U(y)$ among feasible candidates. If no candidate is feasible, the selector records the candidate with the smallest aggregate constraint violation as a diagnostic repair target; the orchestrator does not accept it as the new state. This rule prevents a superficially safe but meaning-altering rewrite from advancing through the loop.

The selector also keeps the dimensions interpretable. The toxicity component receives the largest utility weight after feasibility is established, semantic preservation breaks ties between similarly safe candidates, and fluency rewards polished output without compensating for a threshold failure.

3.5 AMEND Lite

Figure 1 summarizes the implemented loop. Inputs below the safety threshold are returned unchanged only when both the model critic and transparent residual-span checker find no remaining toxic span. Other inputs enter a full first round with three generator roles and three critics.

Lite makes the feedback loop bounded and monotonic. If the selected candidate remains above the toxicity target or contains residual spans, the second round calls only the minimal editor with the critic feedback. A candidate is accepted when it is feasible and either decreases toxicity by at least the minimum step or keeps toxicity from increasing while strictly reducing the residual-span count. The process stops on a safe input, target attainment, lack of improvement, a repeated normalized state, absence of valid candidates, or the two-round limit. These gates retain the multi-role search where diversity is most useful and narrow later computation to residual repair.

4 Experimental Setup

4.1 Data and Deterministic Sampling

Experiments use the English ParaDetox training file (Logacheva et al., 2022b), identified by SHA-256 `6c305890...9c72ebc`. The sampling pipeline filters sources with a transparent harmful-term lexicon, deduplicates by source, assigns a deterministic SHA-256 order, and selects contiguous offsets. The initial ablation skips ten prompt-development examples and evaluates the next 100 inputs. Lite variants are selected on offsets 110–139. The frozen confirmation uses offsets 140–239, which were not used for prompt or control-rule selection. Comparisons within each experiment use aligned inputs and seeds.

4.2 Model and Conditions

All reported model runs use a local 4-bit `qwen3.5-4b-mlx` model through an OpenAI-compatible LM Studio endpoint. Decoding uses temperature 0.2, top-p 0.8, top-k 20, and a 128-token output budget. Experiments use seeds 42, 43, and 44.

The ablation compares three single-role generators, one-round selection over the exact candidates produced by those roles, and legacy AMEND with up to three rounds. Every condition uses the same three critics. The frozen confirmation compares keyword masking, the legacy three-round controller, and AMEND Lite with a full first round and at most one targeted repair.

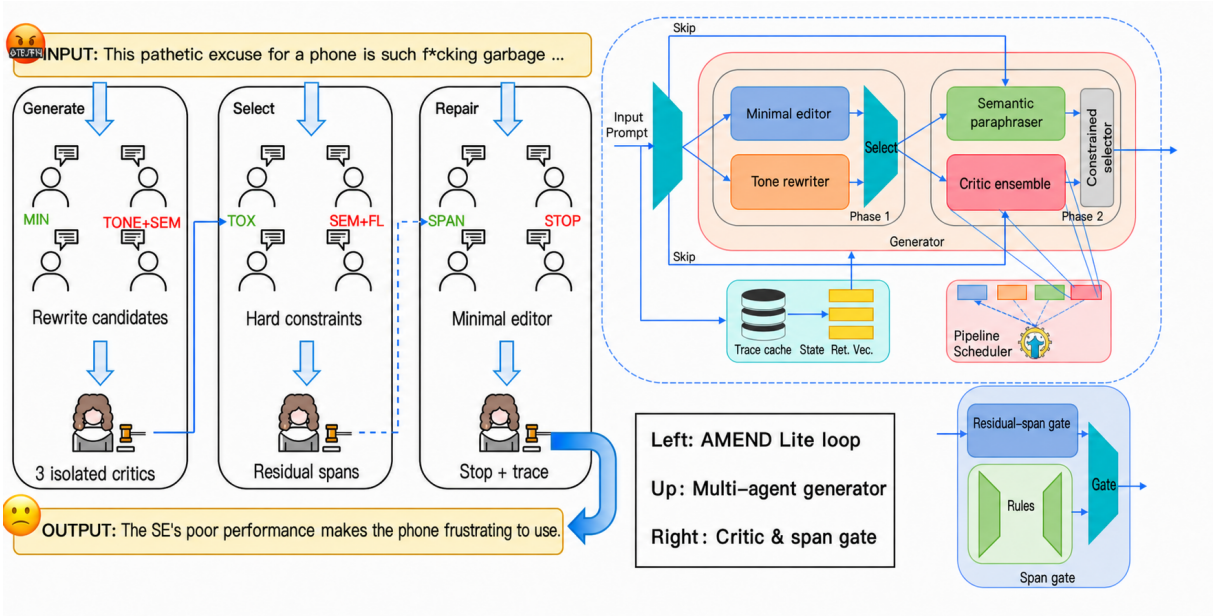

Figure 1: AMEND Lite generates role-diverse candidates, evaluates them with isolated critics, selects under hard preservation and fluency constraints, and invokes one targeted repair when residual spans remain.

4.3 Metrics and Statistical Analysis

We report transparent automatic proxies. **Style (lexicon)** is one when the output contains no complete match from the evaluation lexicon. **Ref-F1** is unigram token F1 against the ParaDetox human neutralization. **Source-F1** is unigram token F1 against the source. **Surface Fluency** penalizes empty strings, abnormal spacing or punctuation, and explicit removal markers. The per-example joint proxy is

$$J_i = \text{Style}_i \times \text{RefF1}_i \times \text{Fluency}_i, \qquad (3)$$

and the reported joint score is the mean of J_i, rather than a product of corpus-level means.

For model conditions, $\pm$ denotes the sample standard deviation of the three seed means. Paired comparisons first average each aligned sample over seeds and then apply 5,000 bootstrap resamples over the 100 sample-level differences. We report the observed difference and percentile 95% interval.

5 Results

5.1 Role Diversity and Constrained Selection

Table 1 reports the initial ablation. One-round selection over the three-role candidate pool reaches a joint proxy of 0.5894 ± 0.0245, compared with 0.5385 ± 0.0080 for the strongest single role. The paired improvement is $+0.0508$, with a 95% bootstrap interval of $[+0.0224, +0.0812]$. The one-round condition also improves over the minimal role by $+0.1270$ and the semantic role by $+0.0752$, with intervals excluding zero in both comparisons.

All three roles contribute selected outputs. Across 276 non-safe first-round selections, the minimal editor is chosen 157 times, the tone rewriter 99 times, and the semantic paraphraser 20 times. Each non-safe input produces 2.36 distinct candidates on average. The distribution shows that conservative editing supplies the largest share of selected rewrites, while broader alternatives provide additional feasible choices.

5.2 Frozen AMEND Lite Confirmation

Table 2 presents the frozen confirmation. AMEND Lite reaches a joint proxy of 0.6597 ± 0.0062, compared with 0.6108 ± 0.0156 for legacy AMEND. Lite increases Style from 0.8167 to 0.8767 while also increasing Ref-F1 from 0.7497 to 0.7582 and Source-F1 from 0.7941 to 0.8067. Surface Fluency remains similar at 0.9870 and 0.9860.

After averaging seeds within each sample, the paired Lite–legacy difference is $+0.0489$, with a 95% interval of $[+0.0212, +0.0788]$. Lite improves over keyword masking by $+0.3836$ with interval $[+0.3286, +0.4360]$, while legacy improves by $+0.3347$ with interval $[+0.2751, +0.3923]$. The three Lite seed means are 0.6661, 0.6536, and 0.6596.

Condition	Style (lex.)	Ref-F1	Source-F1	Surface Fluency	Joint Proxy
Keyword mask	**1.0000**	**0.7512**	—	0.3770	0.2817
Single minimal	0.5867	0.7778	**0.9010**	0.9607	0.4623 ± 0.0122
Single tone	0.7533	0.7102	0.7590	**0.9850**	0.5385 ± 0.0080
Single semantic	0.6967	0.7280	0.7912	0.9820	0.5141 ± 0.0165
Multi-agent, one round	0.8000	0.7384	0.7983	0.9807	0.5894 ± 0.0245
AMEND, up to three rounds	0.8100	0.7395	0.7979	0.9797	$\mathbf{0.5954 \pm 0.0196}$

Table 1: Initial 100-example, three-seed ablation. Joint is the mean of per-example products.

Condition	Style (lex.)	Ref-F1	Source-F1	Surface Fluency	Joint Proxy
Keyword mask	**1.0000**	0.7449	—	0.3710	0.2761
Legacy AMEND	0.8167	0.7497	0.7941	0.9860	0.6108 ± 0.0156
AMEND Lite	0.8767	**0.7582**	**0.8067**	**0.9870**	$\mathbf{0.6597 \pm 0.0062}$

Table 2: Frozen confirmation on a disjoint 100-example subset with seeds 42, 43, and 44.

Measure	Legacy	Lite	Change
Physical requests/item	13.19	9.45	-28.4%
Prompt tokens/item	3466.7	2481.4	-28.4%
Completion tokens/item	654.2	439.7	-32.8%
Raw time/item (s)	54.6	38.0	-30.4%
Rounds/item	1.087	1.007	-7.4%
Candidates/item	2.31	2.10	-9.2%
Est. standalone requests/item	13.19	12.45	-5.6%

Table 3: Execution audit. Physical Lite measurements reuse cached input critiques; the final row restores their three requests per item.

5.3 Iteration and Stopping Behavior

The legacy full controller improves over one-round multi-agent selection by $+0.0061$, with a 95% interval of $[-0.0028, +0.0163]$. Among its 300 seed-level outputs, 243 terminate at the target, 24 use the safe-input path, 11 stop for no improvement, and 22 detect a cycle; none end unresolved at the round limit. These outcomes motivate Lite's allocation of the full role set to the first round and a single repair role to the second.

5.4 Execution Cost

Table 3 separates the physical paired run from an independently deployed request estimate. During the paired run, Lite reuses each input critique from the aligned legacy record. This execution records 9.45 requests per example for Lite and 13.19 for legacy, a 28.4% reduction. Adding the three input-critic requests required by standalone Lite gives an estimate of 12.45 requests per example, approximately 5.6% below legacy. The run completes 600 condition records with no unresolved failures, 6,793 local model requests, and two automatic retries.

6 Analysis

6.1 Complementary Proposal Roles

The selection counts and paired results jointly characterize the role pool. The minimal editor supplies the plurality of accepted candidates and preserves the greatest source overlap, while the tone role is the strongest single-role condition. The semantic paraphraser is selected less often but still produces distinct feasible alternatives. Constraint-guided selection exploits these asymmetric strengths without requiring a fixed role preference. The improvement over every single-role condition follows from choosing among aligned candidates rather than resampling a separate multi-agent condition.

6.2 Why Lite Retains the Main Gain

The ablation assigns most of the measured gain to diverse first-round proposals and constrained selection. Additional open-ended rounds change the joint proxy only slightly. Lite preserves the high-value stage, then uses critic residual spans to narrow the repair operation. Duplicate filtering, monotonic acceptance, and cycle detection further prevent repeated evaluations of equivalent states. This design aligns computation with the observed contribution of each stage: broad search first, targeted correction second.

7 Conclusion

AMEND structures text detoxification as a transparent sequence of diverse proposal generation, isolated multidimensional criticism, constraint-first selection, and bounded repair. The initial ablation shows that selecting from three role-conditioned candidates improves the joint proxy over every

single-role condition. The frozen confirmation shows that AMEND Lite raises the joint proxy from 0.6108 ± 0.0156 to 0.6597 ± 0.0062, with a paired difference of $+0.0489$. The component scores increase together, and the bounded controller retains the first-round gain while directing later computation to residual spans. These findings establish a practical basis for expanding AMEND with independent evaluators, human judgments, and broader model and language coverage.

Limitations

The confirmation covers 100 lexicon-matched English ParaDetox sources and three decoding seeds. It measures a controlled local configuration with a 4-bit Qwen3.5-4B-MLX model; it does not establish behavior for other model families, model scales, domains, or languages. The initial ablation and frozen confirmation use disjoint deterministic offsets, but neither constitutes a full-benchmark comparison.

The reported Style, Ref-F1, Source-F1, Surface Fluency, and Joint values are transparent automatic proxies. Lexicon absence does not capture implicit toxicity, unigram overlap does not fully measure semantic equivalence, and the surface heuristic does not measure human naturalness. The generator and critics share the same base model, creating correlated errors and possible self-preference. The study contains no independent learned toxicity or fluency evaluator and no blinded human evaluation. Accordingly, the results support the evaluated proposal and control mechanisms rather than a state-of-the-art quality claim.

The multi-round difference has a 95% interval crossing zero, so the experiment does not establish a stable gain from additional rounds. The physical Lite execution reuses cached input critiques; its 28.4% request reduction is not an independent deployment estimate. Restoring those requests yields an estimated 5.6% reduction, while uncached token and latency measurements are unavailable.

The repository records exact summaries, aggregation code, dataset hashes, and sampling procedures, but the present checkout does not include the original experiment JSONL files, run manifests, or the ParaDetox data file. The tables are therefore traceable to committed result documents but cannot be regenerated from the public report directory alone. The evaluated model artifact also lacks a committed binary hash and LM Studio version record. DeepSeek experiments, external metrics, human judgments, and cross-lingual evaluation remain outside the evidence presented here.

Ethical Considerations

Text detoxification can support safer communication and moderation workflows, but automated rewriting also changes a speaker's expression. Systems should preserve propositions, identity references, negation, and intended actions, and they should expose the transformed text when authorial control matters. Over-aggressive neutralization may erase legitimate anger, political speech, reclaimed language, or descriptions of abuse. AMEND's semantic constraint and trace improve inspectability but do not replace contextual review.

Deployment should distinguish assistance from covert censorship. A user-facing editor can propose alternatives and let the author choose; an enforcement system requires policy transparency, appeal paths, and evaluation across communities and dialects. Toxic examples should be stored and displayed only where needed for research or auditing, with access controls appropriate to the source data. The report reproduces aggregate measurements rather than unnecessary harmful utterances.

Data Availability

The implementation, test suite, sampling and aggregation code, configuration documentation, and aggregate result reports are available in the AMEND repository. ParaDetox is distributed by its original authors under its own terms. Dataset hashes and deterministic offsets identify the evaluated subsets without redistributing source examples in this report.

Author Contributions

Tianyu Wang conceived the project, designed and implemented AMEND, conducted the experiments, analyzed the results, and prepared the report.

Conflict of Interest

The author declares no conflict of interest.

Funding

This work received no external funding.

Acknowledgments

OpenAI Codex assisted with literature discovery, manuscript drafting, language editing, LaTeX preparation, and consistency checking. The author reviewed the resulting text, verified cited sources and numerical results against the project artifacts, and takes responsibility for the final report.

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 run as a typed trace containing input and output critiques, per-round candidates, feasibility, utility, selected text, status,

round count, and a changed flag. Critic scores are validated to the interval $[0, 1]$. Offline unit tests cover safe-input bypass, English and Chinese control flow, residual-span gating, no-op filtering, targeted second-round repair, constrained selection, joint-metric aggregation, deterministic sampling, and stable group splitting. At report preparation time, all 29 tests passed under Python 3.9.6 and Python 3.13.12, and source compilation passed under Python 3.13.12.

B Paired Comparisons

Comparison	Difference	95% interval
One-round − mask	+0.3077	$[+.2488, +.3656]$
Full − mask	+0.3138	$[+.2552, +.3711]$
One-round − minimal	+0.1270	$[+.0778, +.1789]$
One-round − tone	+0.0508	$[+.0224, +.0812]$
One-round − semantic	+0.0752	$[+.0341, +.1186]$
Full − one-round	+0.0061	$[-.0028, +.0163]$
Lite − legacy	+0.0489	$[+.0212, +.0788]$
Lite − mask	+0.3836	$[+.3286, +.4360]$

Table 4: Sample-aligned joint-proxy comparisons with 5,000 paired bootstrap resamples.