Response to Reviewers

Manuscript: Explainability-Based Token Replacement on LLM-Generated Text
Journal: Journal of Artificial Intelligence Research (JAIR) **Authors:** Hadi Mohammadi, Anastasia Giachanou, Daniel Oberski, Ayoub Bagheri

---

Dear Dr. Kai-Wei Chang and Reviewers,

Thank you for the helpful feedback on our manuscript. We have carefully addressed all concerns raised by the reviewers. Below we describe the changes made in response to each point.

Response to Editor's Summary

1. More detailed analysis of experiment results and ablation study

We added two new subsections in Section 5:

Section 5.x "Analysis of Ensemble Components": This section now provides a detailed breakdown of how each component contributes to the ensemble's performance. We compare the Ensemble against each single model (BERT-base, DistilBERT, XLM-RoBERTa) across all language-domain combinations. We show that the Ensemble achieves an average F1 of 0.86 in English compared to 0.78 for BERT-base, and 0.78 in Dutch compared to only 0.55 for BERT-base. We also analyze the complementary strengths of the frozen and fresh branches, and discuss robustness under adversarial conditions.

Section 5.x "Sensitivity Analysis": This section examines the trade-off between textual fidelity and evasion effectiveness. We analyze how different token selection strategies (SHAP, LIME, Random) affect detection performance and discuss the relationship between the number of tokens modified and evasion success.

2. Better positioning in literature and comparisons with related methods

We expanded the Related Work section to include recent papers that were previously missing:

- David & Gervais (2025) "AuthorMist: Evading AI Text Detectors with Reinforcement Learning"
- Amara et al. (2024) "SyntaxShap: Syntax-aware Explainability Method for Text Generation"
- Moraliyage et al. (2025) "Explainable AI with Integrated Gradients for Detection of Adversarial Attacks"

We now discuss how our approach relates to these works and clarify our contributions.

3. Human evaluation

We conducted a human evaluation study with 100 text samples and 2 evaluators. Key findings:

- Human evaluators achieved only 47.5% detection accuracy overall
- Rewritten texts that evaded machine detection also fooled humans (only 34.5% correctly identified)
- All conditions maintained high fluency (3.85–4.47) and coherence (3.58–4.65) ratings
- We report inter-rater reliability using Cohen's kappa

This finding is now also mentioned in the Abstract.

Response to Reviewer A

Concern: Gains over baseline are marginal (1%)

We added statistical significance discussion in Section 4.1. We also clarify that the Ensemble's main advantage is consistency across diverse conditions, not just raw accuracy. When processing thousands of texts daily, even small percentage improvements lead to big differences in practice.

Concern: Test set details unclear

We clarified the data split in Section 3 and Section 4.1 (10% test set, ~972 samples total, balanced across languages).

Concern: Missing cost analysis

We added computational cost discussion in the Limitations section.

Concern: Missing comparison with watermarking

We discuss watermarking limitations in the Introduction and acknowledge this as future work.

Concern: Only SHAP and LIME explored

We now cite SyntaxShap and Integrated Gradients as alternative XAI methods and note this as a direction for future work.

Concern: Impact of data augmentation not shown

We address this in the new Analysis of Ensemble Components section, comparing frozen (AuTexTification) and fresh (augmented CLIN33) branches.

Concern: Dense tables need visualization

We improved Figure 2 (UpSet plots) to better visualize model overlap patterns.

Response to Reviewer C

Concern: Incremental novelty

We strengthened the Contributions paragraph in Discussion to clarify our novel dual perspective (attack and defense in unified framework) and the frozen-plus-fresh architecture.

Concern: Missing statistical rigor

We acknowledge this limitation and added discussion of significance in Section 4.1.

Concern: Missing ablations

We added the "Analysis of Ensemble Components" subsection as a proxy ablation study. We acknowledge that full leave-one-out ablation would require retraining and note this for future work.

Concern: Missing related work (AuthorMist, SyntaxShap, Moraliyage)

All three papers are now cited and discussed in Related Work.

Concern: Need for failure case analysis

We added discussion of why the Ensemble remains robust in the Analysis of Ensemble Components section, explaining the redundant detection pathways.

Concern: Questions about generality beyond CLIN33

We acknowledge this limitation and mention it in the Limitations section, noting the need for evaluation on larger and more diverse datasets.

Concern: No human assessment of rewriting quality

We added a Human Evaluation section with 100 samples rated for fluency, coherence, and detection by 2 evaluators.

Concern: Figure 2 interpretation cursory

We expanded the discussion of overlap patterns in Section 5, explaining why certain samples are robustly detected.

Concern: Watermarking not systematically studied

We acknowledge this as future work in the Limitations section.

Summary of Changes

1. **New Section: Analysis of Ensemble Components** – Detailed breakdown of component contributions
2. **New Section: Sensitivity Analysis** – Trade-off analysis between fidelity and evasion
3. **New Section: Human Evaluation** – 100 samples, 2 evaluators, fluency/coherence/detection ratings
4. **Expanded Related Work** – Added AuthorMist, SyntaxShap, Moraliyage references
5. **Updated Abstract** – Now mentions human evaluation results (47.5% accuracy)
6. **Expanded Limitations** – Addresses dataset size, computational cost, generalization concerns
7. **Clarified Contributions** – Stronger positioning of novel aspects

We believe these revisions address all major concerns raised by the reviewers. We thank the reviewers for their constructive feedback, which has significantly improved the paper.

Sincerely,

Hadi Mohammadi, Anastasia Giachanou, Daniel Oberski, and Ayoub Bagheri