Explanation of Revisions: The Feasibility of Topic-Based Watermarking on Academic Peer Reviews

Overview of Changes: This resubmission addresses the concerns raised by reviewers in the previous round, particularly regarding the paper's scope, experimental presentation, and justifications. We have made substantial revisions to better position this work as a focused application study while strengthening the empirical evaluation and analysis.

Major Revisions

1. Paper Scope and Positioning: Revised the paper from a long-form submission to a short paper format.
   Rationale: Following reviewer feedback that noted this work is primarily an application of existing TBW methodology to a new domain, we have repositioned the paper to better reflect its contribution as a timely and important application study. This format better aligns with the nature of our empirical evaluation and domain-specific insights.
   The revised scope allows us to focus on the key findings and practical implications while maintaining the comprehensive experimental evaluation that demonstrates TBW's viability in peer review contexts.
2. Baseline Method Comparisons: Moved comparisons with KGW and SynthID watermarking methods from the appendix to the main text.
   Rationale: Multiple reviewers noted that baseline comparisons were essential to the paper and should not be relegated to appendix material. This was particularly emphasized in the meta-review.
   Specific Changes:
   - Integrated KGW and SynthID results in Section 4.1 (Generation Quality).
   - Integrated KGW and SynthID results in Section 4.2 (Robustness to Paraphrasing).
   - Included direct performance comparisons in main results.
   - Provided clearer discussion of why TBW outperforms these baselines in main text.
3. Few-Shot Performance Analysis: Enhanced explanation and analysis of the observed performance degradation in few-shot configurations.
   Rationale: Reviewers requested deeper analysis of why few-shot performance drops.
   Specific Changes:
   - Added detailed explanation in Section 4.2 of topic mismatch between few-shot exemplars and target papers.
   - Included discussion of how this mismatch weakens topic alignment and reduces detectability post-paraphrasing.
   - Provided concrete reasoning for why this limitation can be mitigated through better exemplar selection.

    - Enhanced the discussion of this finding's implications for practical deployment.

4. Strengthened Rationale for TBW Selection: Section 3.3 provides stronger justification for why TBW is well-suited for peer review applications and is a prominent section in the now short paper.
   Rationale: Reviewers questioned why TBW was selected over other watermarking approaches and requested deeper justification for the method's advantages.
   Specific Changes:
    - Enhanced discussion of how the topic-matching assumption naturally holds in peer review contexts.
    - Strengthened explanation of TBW's robustness advantages in realistic threat models.
    - Clarified how peer review constraints (topical consistency requirements) align with TBW's design principles.

Minor Revisions

- Improved clarity in experimental setup descriptions.
- Enhanced discussion of practical deployment considerations.
- Refined the presentation of results.

Maintained Strengths

We preserved the comprehensive empirical evaluations that reviewers praised, including:

- Multi-configuration analysis (base, few-shot, fine-tuned).
- Robustness evaluation under realistic paraphrasing attacks.
- Quality preservation analysis using multiple metrics.
- Classifier-based attribution assessment.

Conclusion

These revisions directly address the primary concerns raised by reviewers while maintaining the core strengths of our empirical evaluation. The repositioned paper better reflects its contribution as a thorough application study that provides valuable insights for the academic community grappling with LLM usage in peer review processes. The enhanced baseline comparisons and theoretical justification strengthen the work's impact while the improved scope positioning sets appropriate expectations for the contribution type.