Explanation of Revisions

This document outlines the revisions made in the resubmitted version of our paper, "POCO: Low Cost Post-OCR Correction Dataset Construction via Character-Level Probabilistic Simulation", compared to the previous submission.

1. Main Contribution Shift and Research Area Update
We revised the framing of our paper to better align with its primary contribution. While the earlier version emphasized OCR correction performance, the revised version highlights our synthetic dataset construction method (POCO) as the core contribution. Accordingly, we updated the research area from application-centric to Resources and Evaluation.

2. Expanded Experimental Analysis
To strengthen our evaluation:

- We added an analysis of OCR error types across languages and OCR models (Appendix B, Table 4).
- We investigated the cause of large gains in Korean PaddleOCR results, attributing it to excessive deletion errors in punctuation and whitespace.
- We visualized the length distribution of OCR and corrected sentences with and without spaces for better insight into spacing-related errors (Appendix D).

3. Additional Model Comparisons
In response to concerns about model choice, we extended our simulator experiments:

- EfficientNet, ViT, and MobileNet were trained in addition to ResNet34.
- We compared training loss (Figure 7) and per-epoch training speed (Appendix A, Table 3), finding that EfficientNet provides a good trade-off between accuracy and speed.

4. Prompts and LLM Behavior
We expanded our analysis of GPT-4o:

- We added the exact prompts used during LLM-based correction (Appendix C, Figure 8).
- We discussed why GPT-4o underperformed, highlighting its tendency to alter structure rather than correct characters.

5. Clarified Limitations Section
We revised and expanded the Limitations section to clearly discuss:

- Character-level modeling constraints
- Fixed error rate assumptions
- Limited language coverage (Korean and Chinese only)
- Lack of contextual or downstream task evaluation

6. Other Improvements
We revised the abstract and key sections for clarity and brevity.
We plan to release our code, which is now stated clearly at the end of the abstract.
Figures and tables were reorganized and captioned more precisely.