

CIKM Short Paper: Knowing When to Correct

Venue: ACM International Conference on Information and Knowledge Management (CIKM)

Authors: Stergios Konstantinidis, Hayman Lotfy, Michalis Vlachos

Affiliation: Department of Information Systems (DESI), HEC Lausanne

GitHub Public: [CIKM_public](#)

Overleaf Project: <https://git.overleaf.com/6a1d78e803cbdf7def32a839>

Core Codebase: [/Users/stergios/Documents/GitHub/Optimizing-LLM-based-OCR-Corrections](#)

1. Problem Statement

At multi-million page scale, LLM-based OCR correction is constrained by cost:

- Full LLM correction costs hundreds of thousands of dollars on national archival collections.
- Many archival passages are already accurate; invoking LLMs on them introduces latency, cost, and risk of hallucinations.

Core Research Questions:

- **RQ1:** How closely does selective routing match the theoretical oracle ceiling, and does it outperform naive confidence thresholding?
- **RQ2:** How well does the regression-estimated improvement $\hat{\Delta}_i$ correlate with the true realized improvement Δ_i ?
- **RQ3:** What is the actual operational cost saved when routing only segments predicted to benefit across various frontier LLMs?

Problem Visual Illustration

FABRIQUE DE CHAPEAUX DE PAILLE DE OBERSON-MULLER
RUE DE BOURG, 16, MAISON DU LION D'OR.
7. Un choix de chapeaux pour Dames et Messieurs de tous gen.



Raw OCR

FABRIQUE DESCHAPBAUX DE PAILLE DE OBRRSON-
MULLER\nRUE DE BourG, 16,
MAISON pu Lion D'oR.\n7.Un choix de chapeaux pour
Dames et Messieurs de tons gen.



OCR + LLM Corrected

FABRIQUE DE CHAPEAUX DE PAILLE DE OBERSON-
MULLER\nRUE DE BOURG, 16, MAISON DU LION
D'OR.\n7.Un choix de chapeaux pour Dames et
Messieurs de tous gen.

Figure 1: Typical OCR degradation in historical periodicals requiring cost-aware routing.

Revision #7

Created 2026-09-09 16:20:19 UTC by Stergios

Updated 2026-09-17 12:21:53 UTC by Stergios