Presentation slides

□□ DocEng 2026 Conference Presentation Deck

Official 10-slide presentation delivered at ACM DocEng 2026 (28.08.2026).

□□ Download PDF □□ Download PPTX

Slide Deck & Presentation Breakdown

Slide 1: Title & Authors

Slide 01 / 10

Cost-Aware Human-LLM Collaboration for Post-OCR Corrections in Swiss Historical Newspapers

DocEng26

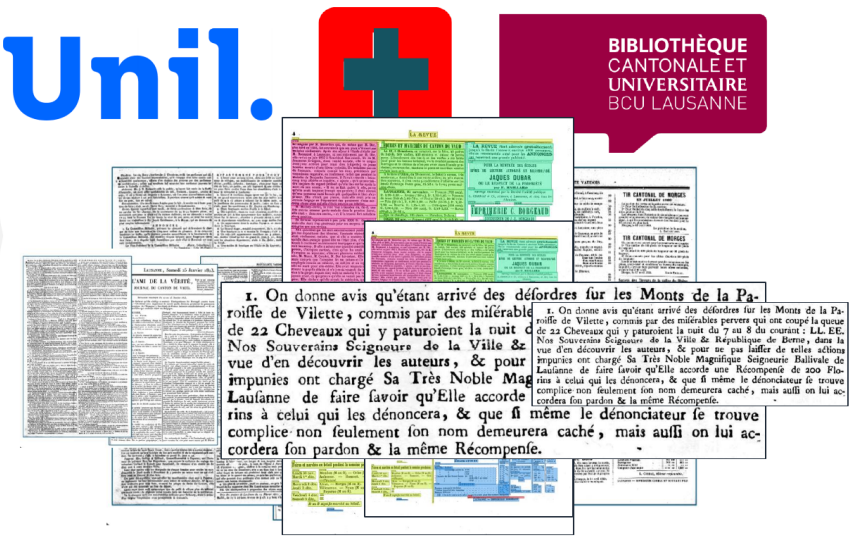
28.08.2026 | Stergios Konstantinidis | Hayman Lotfy | Prof. Michalis Vlachos

Speaker Note: Cost-Aware Human-LLM Collaboration for Post-OCR Corrections in Swiss Historical Newspapers (DocEng 2026, 28.08.2026). Presented by Stergios Konstantinidis, Hayman Lotfy, Prof. Michalis Vlachos.

Motivation | We are working on RAG systems for information retrieval over historical archives...

Meet our data:

- **Thousands of newspaper issues spanning over 300 years**
- Millions of pages
- 12 tb of raw images –> 1 image = 1 page

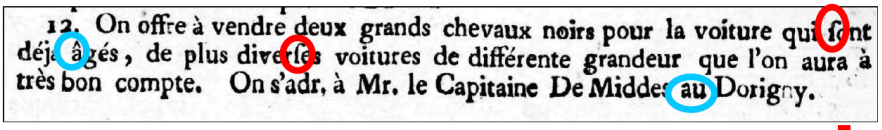


Speaker Note: Building RAG systems for historical archives across 300+ years of newspapers, millions of pages, and 12 TB of raw high-resolution scan imagery.

Slide 3: Why Document Cleaning is Not

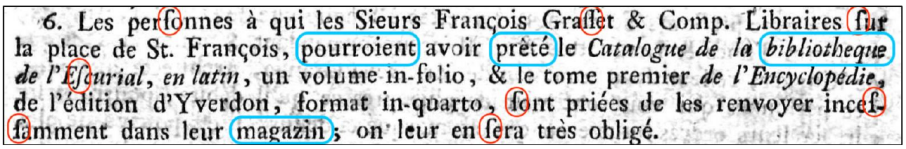
Motivation | Why document cleaning is not always straightforward
| An OCR text corrector use-case

- Post-OCR text correction is an essential step to improve the quality of our corpus
- Traditional correction tools often neglect linguistic shifts



12. On offre à vendre deux grands chevaux noirs pour la voiture qui ſont déjà âgés, de plus diverſes voitures de différente grandeur que l'on aura à très bon compte. On s'adr. à Mr. le Capitaine De Middes au Dorigny.

12. On offre à vendre deux grands chevaux noirs pour la voiture qui sont déjà agés, de plus diverses voitures de différente gradeur que l'on aura à très bon compte. On s'adr. à Mr le Capitaine De Middes a Dorigny.



6. Les perſonnes à qui les Sieurs François Graſſet & Comp. Libraires ſur la place de St. François, pourroient avoir prêté le *Catalogue de la bibliotheque de l'Eſcurial, en latin*, un volume in-folio, & le tome premier *de l'Encyclopédie*, de l'édition d'Yverdon, format in-quarto, ſont priées de les renvoyer inceſſamment dans leur magazin; on leur en ſera très obligé.

6. Les perſonnes à qui les Sieurs François Grasſet & Comp. Libraires ſur la place de St. François, pourraient avoir prêté le Catalogue de la bibliothèque de l'Escurial, en latin, un volume in-folio, & le tome premier de l'Encyclopedie, de l'édition d'Yverdon, format in quarto, ſont priées de les renvoyer inceſſamment dans leur magasin; on leur en ſera très obligés

page 03

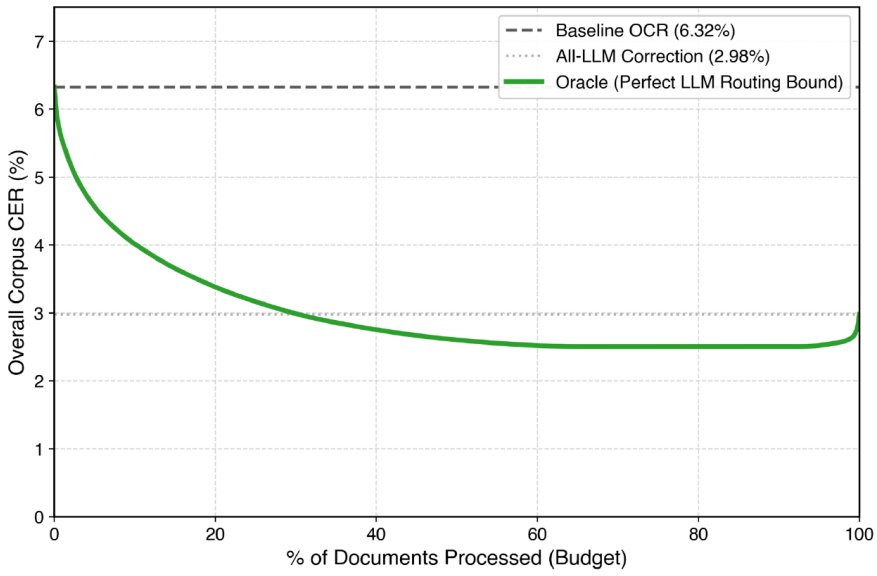
Speaker Note: Post-OCR correction is essential for downstream retrieval, but traditional tools neglect historical linguistic shifts and archaic orthography.

Slide 4: Heterogeneous Segment Utility



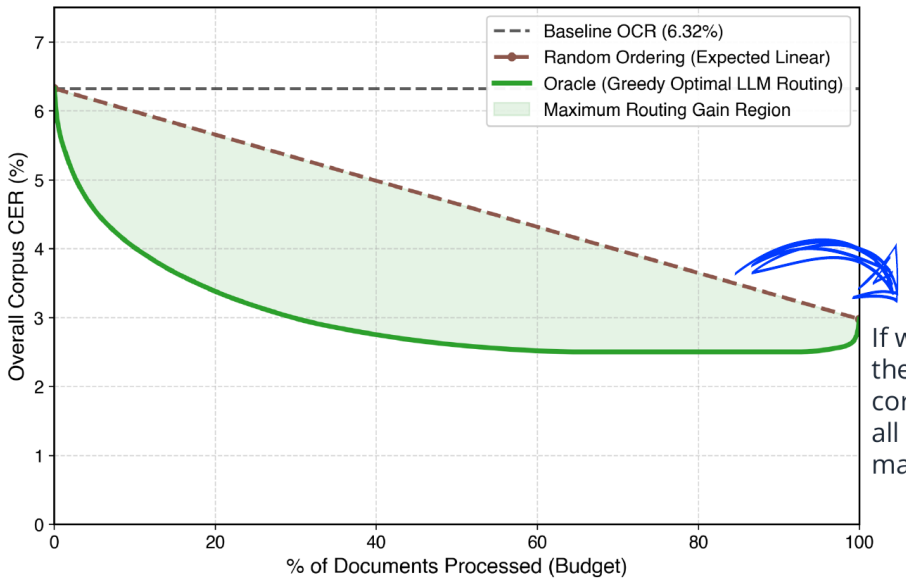
Speaker Note: Different segments benefit very differently: Segment D gains +2.0% CER improvement with zero overcorrections, while Segment B suffers -3.0% CER degradation due to 10 overcorrections!

Motivation | Assume we know in advance the documents that would benefit most by being corrected:



Speaker Note: Theoretical premise: What if we knew in advance which archival documents would benefit most from LLM correction?

Motivation | Assume we know in advance the documents that would benefit most by being corrected: In reality we don't know the right order!



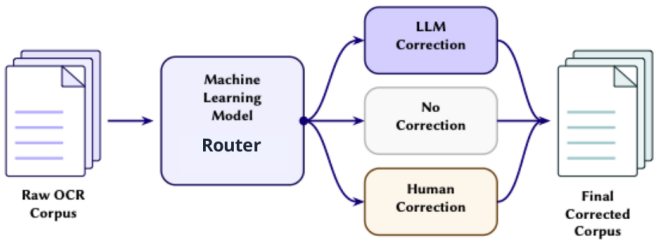
If we don't know the right order to correct, correcting all documents makes sense!

Speaker Note: In reality, ground truth is unavailable. Naively correcting all documents produces unpredictable degradation and wastes massive token budgets.

Slide 7: Our Contribution - Intelligent Routing Architecture

Slide 07 / 10

Implementation:

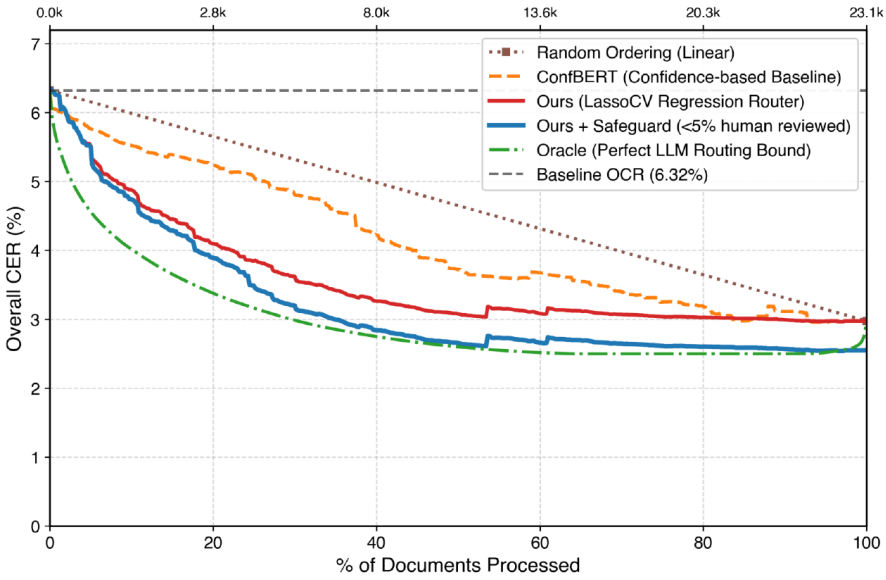


- Router that evaluates the potential improvement that may be offered by LLM corrections.
  - If LLM correction is sufficient -> route to LLM
  - Elif too complex -> human review
  - Else leave the segment alone

Speaker Note: A machine-learned Router evaluating expected CER benefit: If LLM is sufficient -> route to LLM; if degradation is too severe -> route to human review (<5%); otherwise leave the segment untouched.

Slide 8: Experimental Evaluation & Benchmark Curves

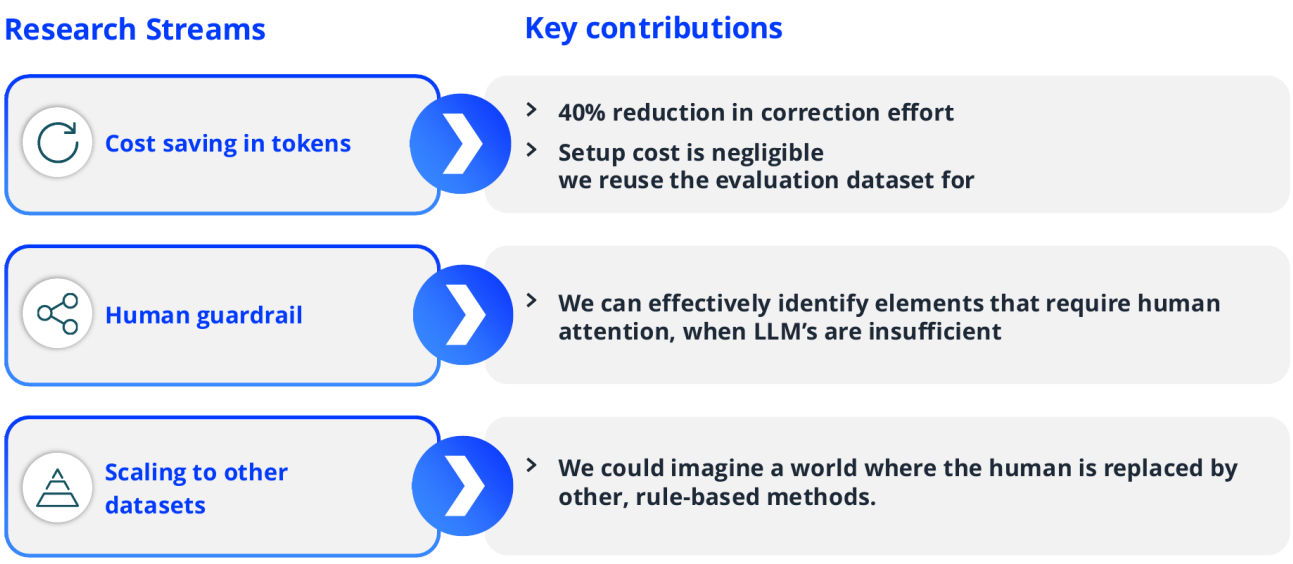
Slide 08 / 10



Speaker Note: Empirical results comparing baseline OCR, All-LLM, ConfBERT, and our Cost-Aware Routing framework across historical Swiss periodicals.

Slide 9: Key Takeaways & Research Streams

Slide 09 / 10



Speaker Note: Key contributions: 40% reduction in correction effort, negligible setup cost, and an effective human guardrail identifying segments requiring expert archivist attention.



Speaker Note: Wrap-up, open questions, and research collaboration contact information at UNIL-DESI.

Revision #4
Created 2026-09-17 11:44:37 UTC by Stergios
Updated 2026-09-17 12:21:52 UTC by Stergios