

The background of the slide is a scenic landscape featuring a large body of water, likely Lake Geneva, with a range of mountains in the distance. The entire image is covered with a semi-transparent blue filter. In the foreground, there are green trees and a modern, circular building with a tiered, amphitheater-like structure.

Unil.

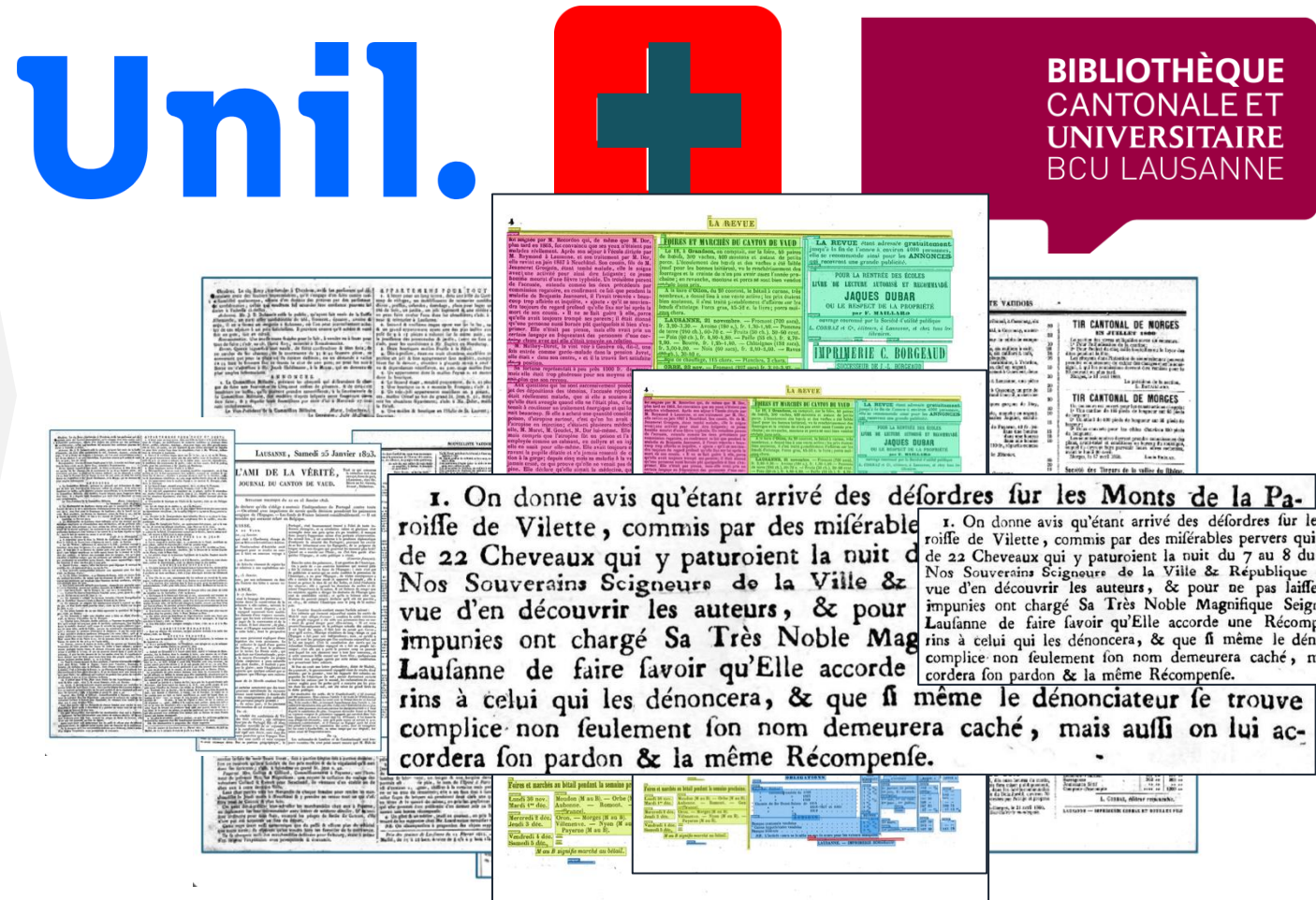
Cost-Aware Human-LLM Collaboration for Post-OCR Corrections in Swiss Historical Newspapers

DocEng26

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



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

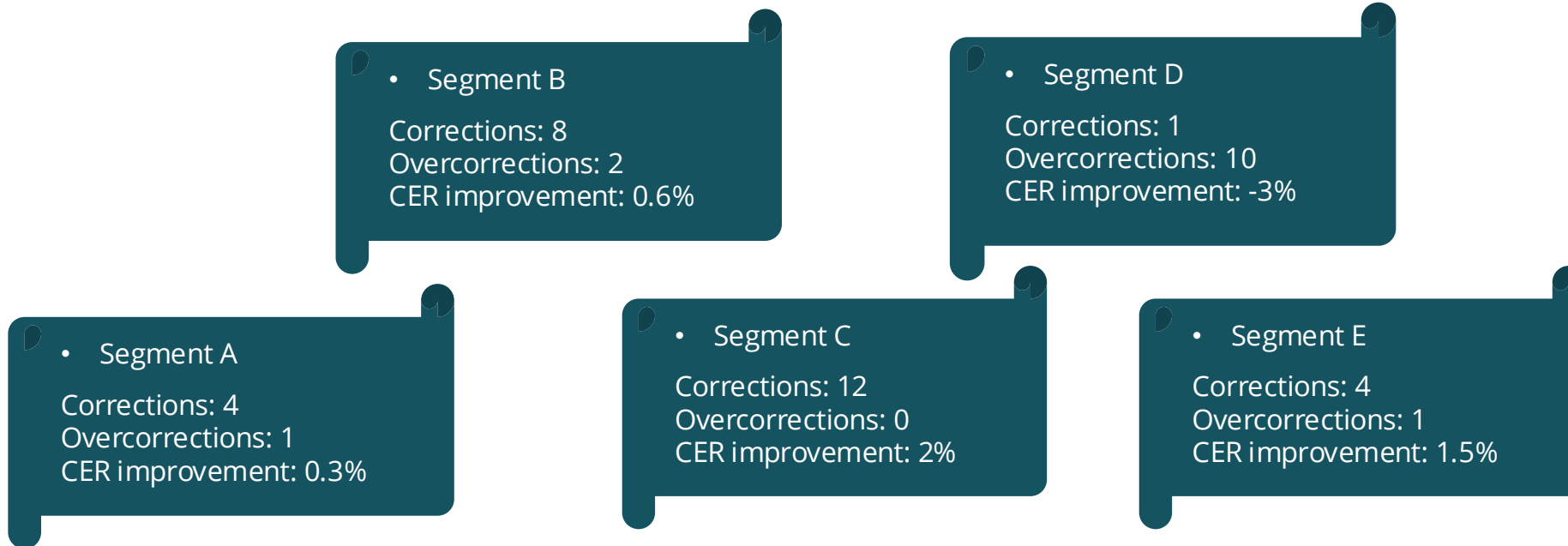
12. On offre à vendre deux grands chevaux noirs pour la voiture qui sont déjà âgés, de plus diverses voitures de différente grandeur que l'on aura à très bon compte. On s'adr. à Mr. le Capitaine De Middelau Dorigny.

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 Middel a Dorigny.

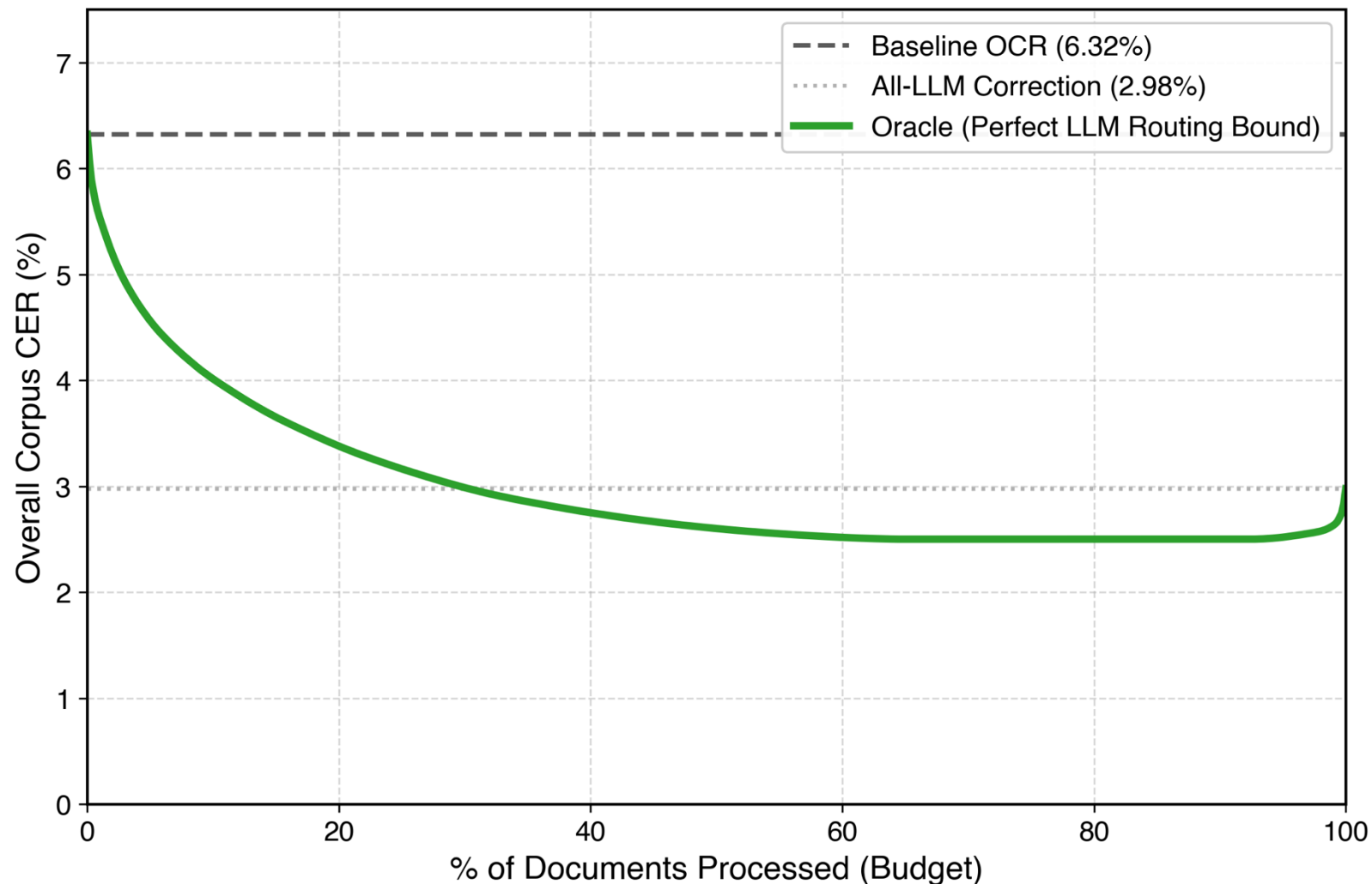
6. Les personnes à qui les Sieurs François Grasset & Comp. Libraires sur la place de St. François, pourroient avoir prêté le Catalogue de la bibliothèque de l'Escurial, en latin, un volume in-folio, & le tome premier de l'Encyclopédie, de l'édition d'Yverdon, format in-quarto, sont priées de les renvoyer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

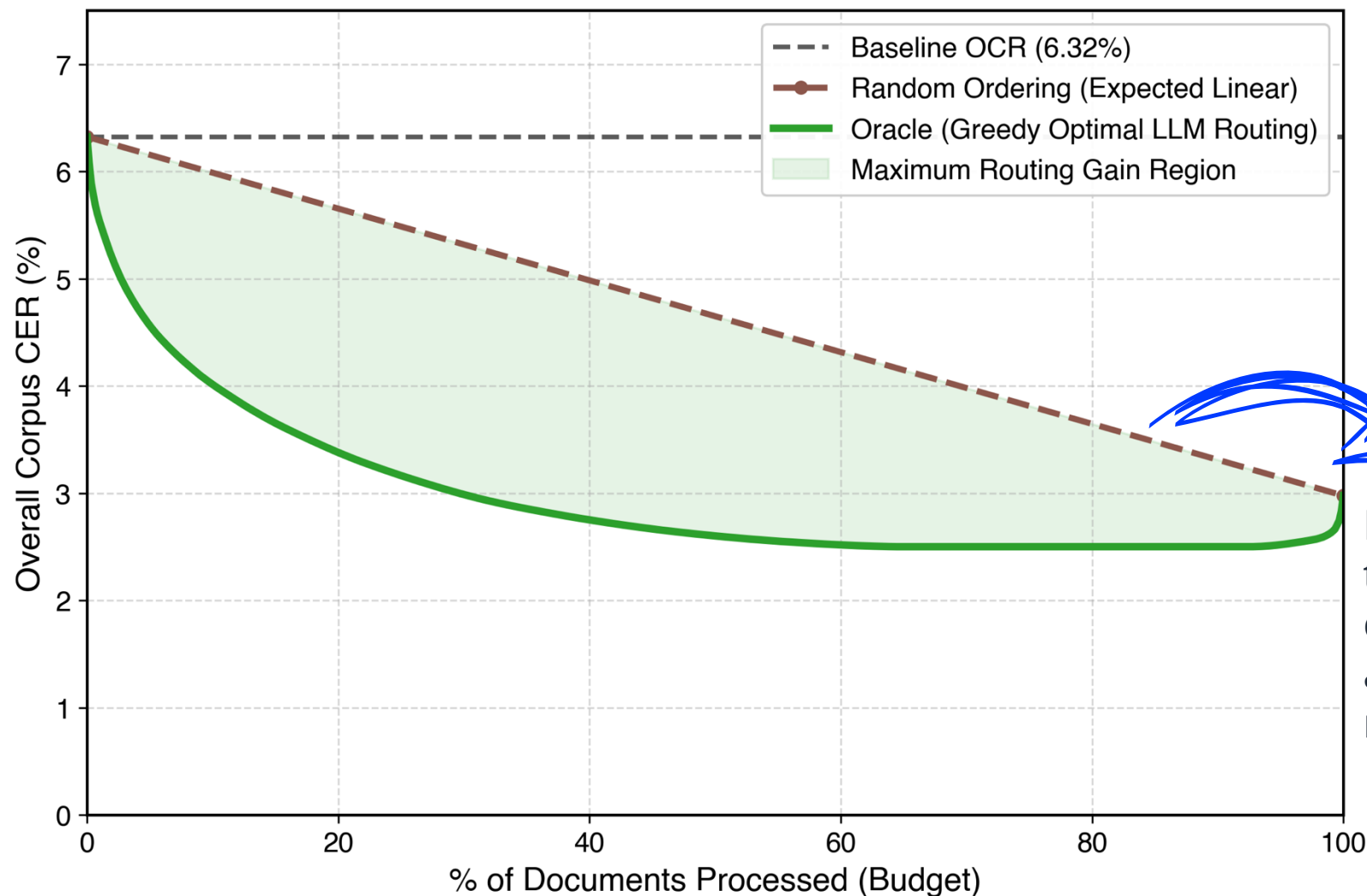
Motivation | Different segments benefit differently by text corrections



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



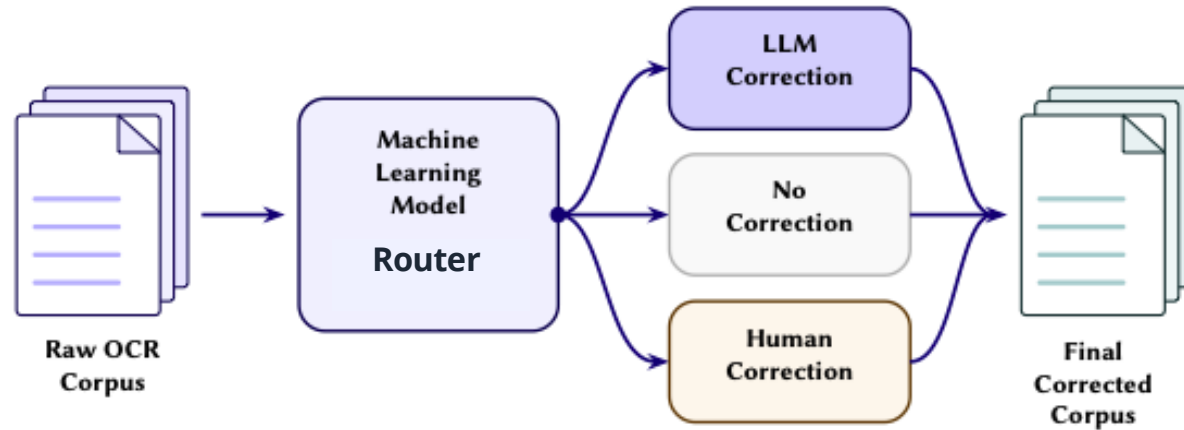
Motivation | Assume we know in advance the documents that would benefit most by being corrected: In reality we don't know the right order!



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

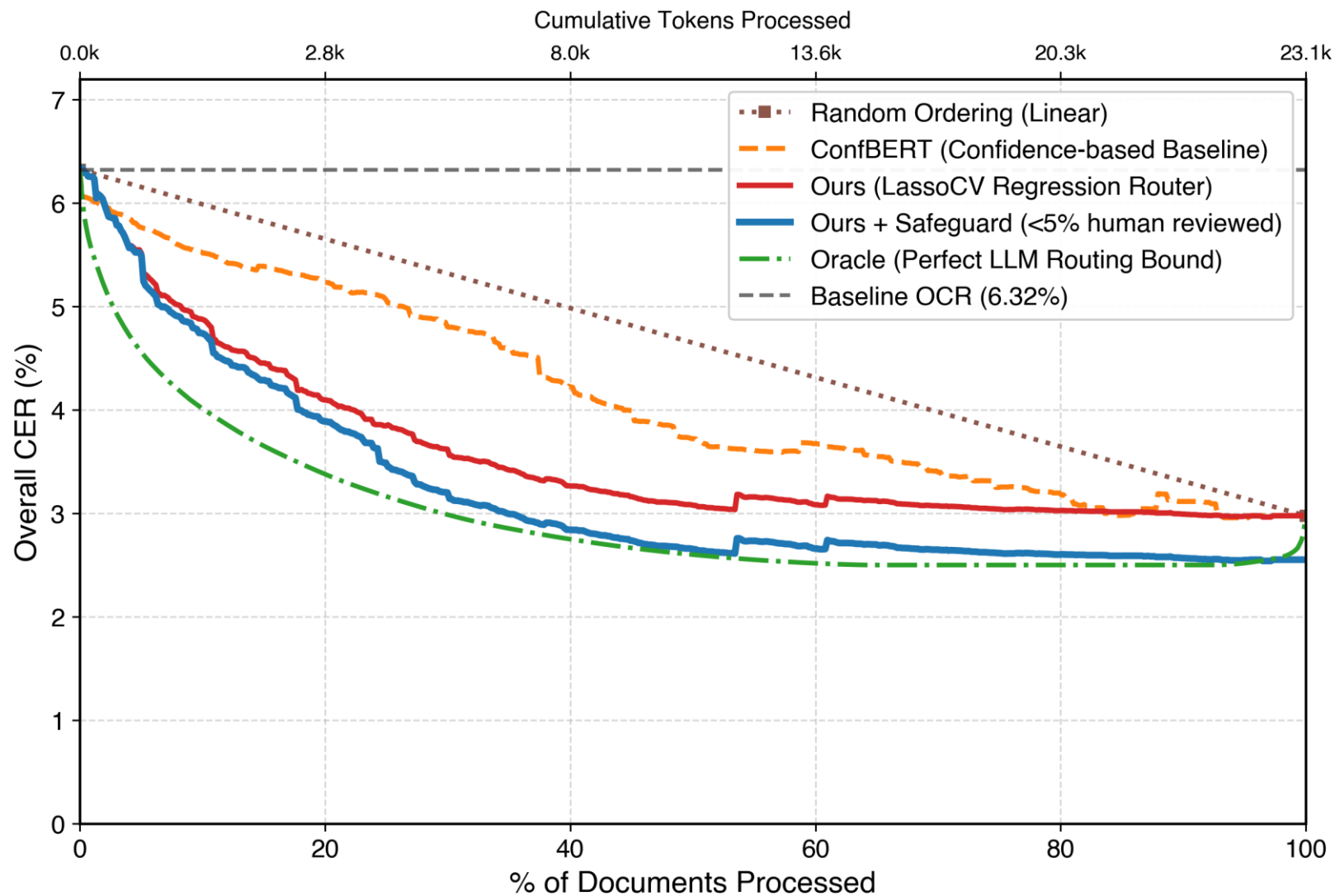
Our contribution | What if we knew the right order in which to correct the documents?

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

Results | How well does our model perform?



Summary | Key takeaways from our paper

Research Streams

Key contributions



Cost saving in tokens



- > 40% reduction in correction effort
- > Setup cost is negligible
we reuse the evaluation dataset for



Human guardrail



- > We can effectively identify elements that require human attention, when LLM's are insufficient



Scaling to other datasets



- > We could imagine a world where the human is replaced by other, rule-based methods.

THANK YOU!



Get in touch!