Dataset & Benchmark Results

Experimental Benchmark & Evaluation

1. Archival Corpus

Evaluated across **609 text segments** spanning **45 digitized pages** from 9 Swiss French-language historical periodicals (Canton of Vaud, 1733–1945):

Periodical	Era	Issues	Pages	Segments	Characteristics
La Revue	1875–1945	4	5	139	19th/20th century serif, multi-column
Feuille d'Avis	1762–1841	4	6	118	Antique typography, long 's' ligatures
Mercure Suisse	1733–1748	3	5	92	18th-century early modern French
Gazette de Lausanne	1804–1920	5	8	145	High density, complex layout
Others (Tribune, etc.)	1880–1930	5	21	115	Mixed commercial & political press

2. Key Benchmark Results

Error Rate Comparison (Tesseract OCR Base):

- **Raw OCR Baseline:** *CER = 6.30\%, WER = 14.20\%*
- **All-LLM Baseline (GPT-4o / Gemini 1.5 Flash):** *CER = 2.92\%* (Cost: 100% token budget)
- **Selective Pre-Routing (Ours):** *CER = 3.12\%* (Bypasses **40%** of all segments!)
- **Collaborative Safeguard (<5% Human Budget): *CER = 2.49\%*** (**14% relative reduction** over All-LLM)

Comparison with Routing Baselines:

1. **Random Routing:** Weakest; frequently sends clean text and degrades it.
2. **OCR Confidence (Thresholding):** Underperforms because Tesseract confidence scores poorly correlate with semantic error in archaic fonts.
3. **ConfBERT:** Good detection but computationally heavy.
4. **Our Regression Router + Safeguard:** Consistently establishes the Pareto frontier across all budget limits.

Experimental Evaluation & Visual Benchmark Plots

Historical Newspaper Segmentation Sample

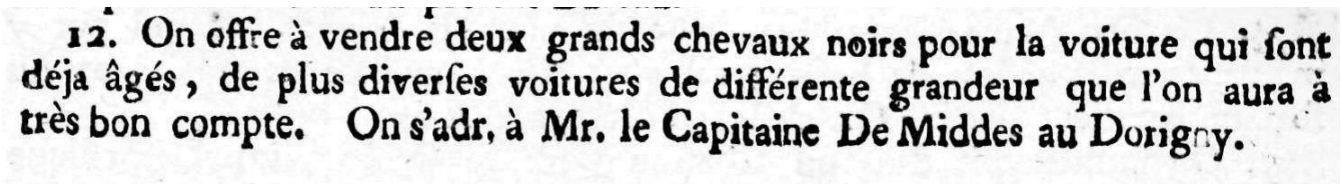

12. On offre à vendre deux grands chevaux noirs pour la voiture qui ſont déja âgés, de plus diverſes voitures de différente grandeur que l'on aura à très bon compte. On s'adr. à Mr. le Capitaine De Middes au Dorigny.

Figure 1: Representative historical archival newspaper segment illustrating layout artifacts and OCR degradation.

CER Reduction Across Correction Strategies

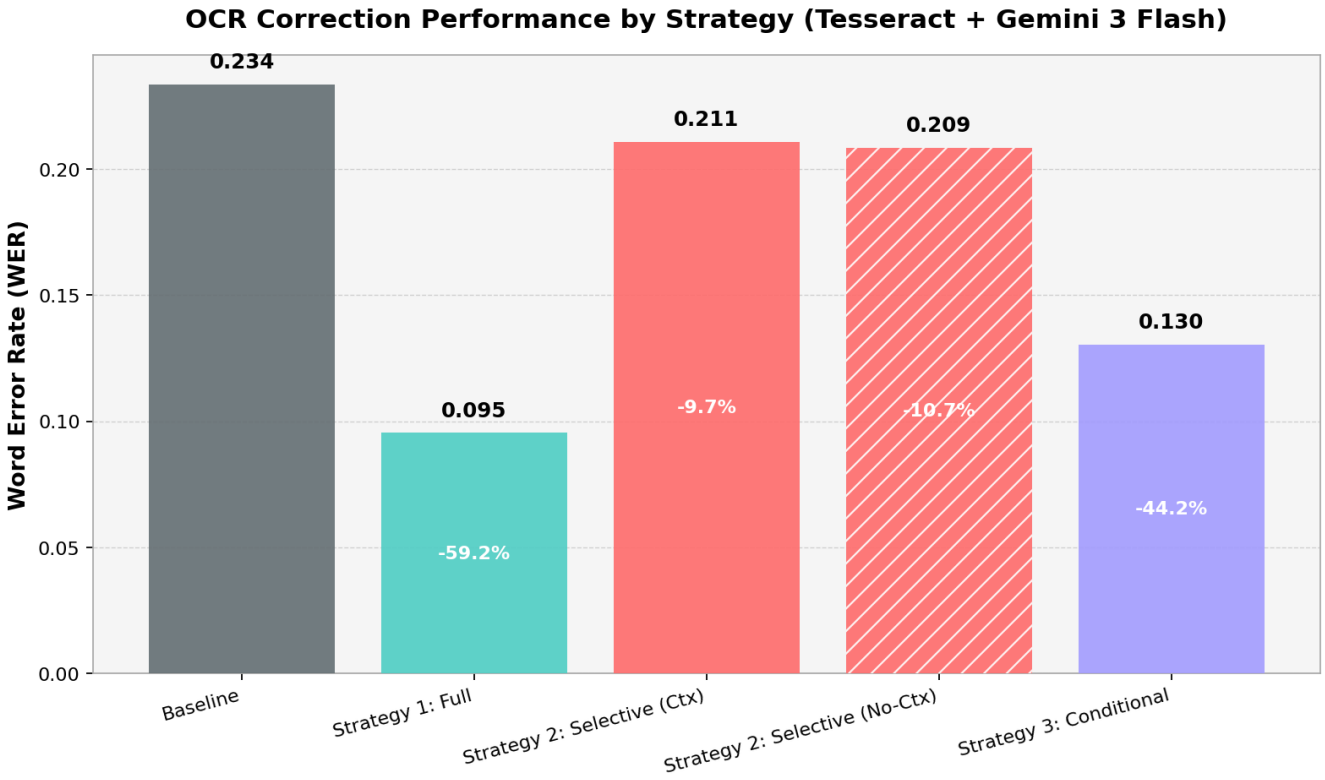


Figure 2: Character Error Rate (CER) reduction across baseline OCR, All-LLM, ConfBERT, and Collaborative Routing.

OCR Engine Improvement Comparison

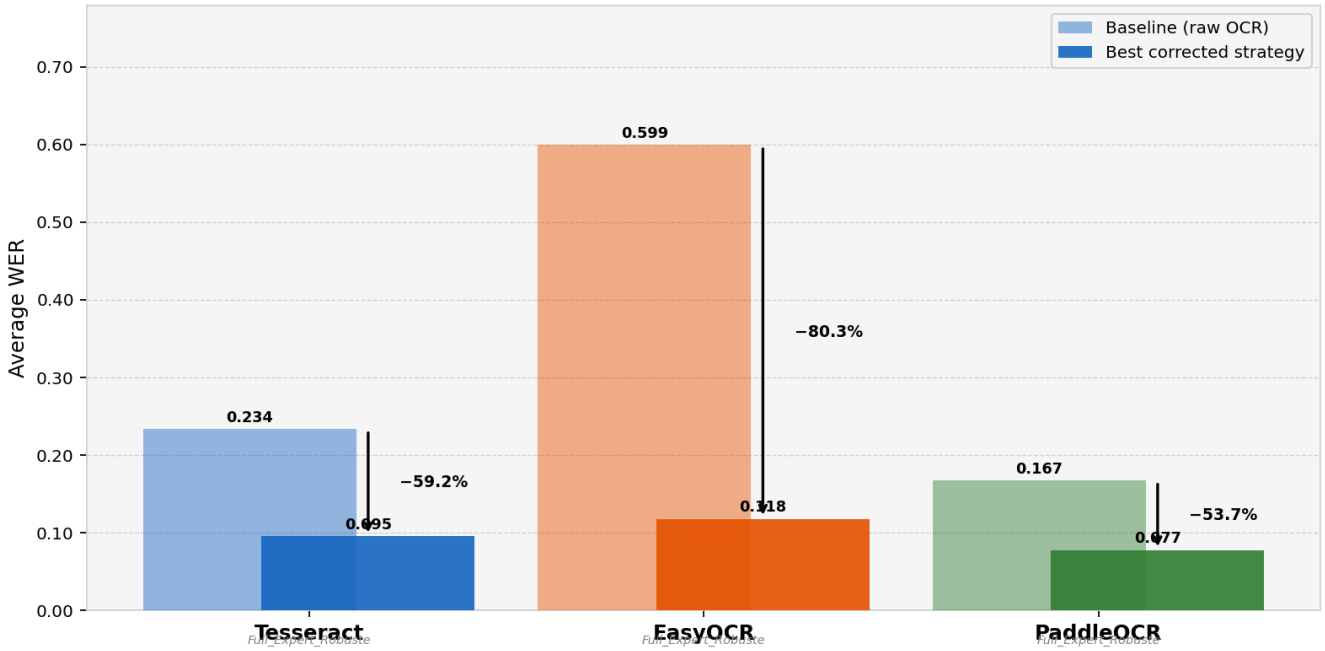


Figure 3: Relative error reduction across historical OCR engines (Tesseract, Abbyy, and modern HTR).

Strategy Performance Heatmap

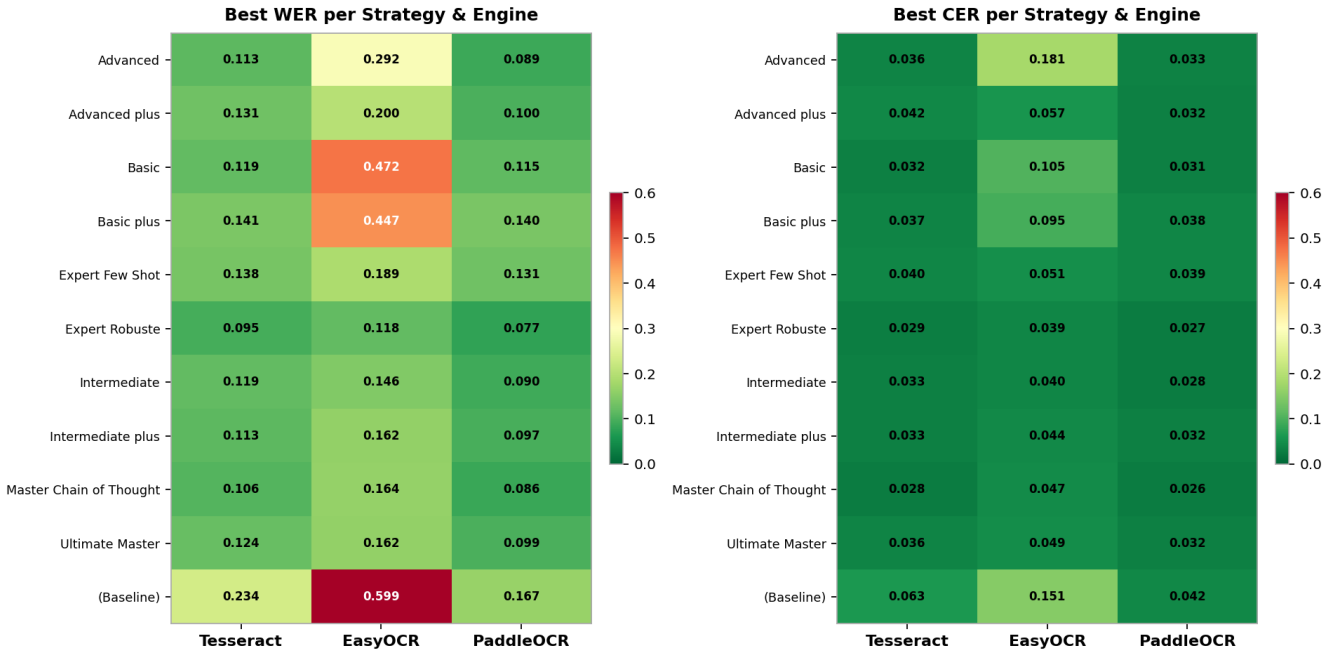


Figure 4: Comprehensive performance heatmap comparing LLM prompt strategies across historical periodicals.

Revision #5
Created 2026-09-09 16:20:17 UTC by Stergios
Updated 2026-09-17 12:21:52 UTC by Stergios