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

Stergios Konstantinidis
University of Lausanne
Lausanne, Switzerland
stergios@unil.ch

Hayman Lotfy
University of Lausanne
Lausanne, Switzerland
hayman.lotfy@unil.ch

Michalis Vlachos
University of Lausanne
Lausanne, Switzerland
michalis.vlachos@unil.ch

Abstract

OCR transcription errors in historical archives often hinder digital search and retrieval. While Large Language Models (LLMs) can correct many of these errors, applying them indiscriminately is costly and may negatively affect already-clean text. We propose a three-tier collaboration framework that routes each text segment to one of: (1) No Correction, (2) LLM Correction, or (3) Human Correction. We introduce a regression-guided routing approach that prioritizes segments by predicted CER improvement, paired with a safeguard layer that detects harmful LLM corrections and routes uncertain segments to human review. With only <5% of the corpus reviewed by human experts, our safeguard achieves a 14% relative reduction over the All-LLM baseline, and substantially outperforms standard confidence-based approaches. By dynamically routing degraded segments to humans and fixable errors to the LLM, the collaborative framework outperforms either corrector in isolation.

Keywords

Large Language Models, Historical Archives, Document Analysis

ACM Reference Format:
Stergios Konstantinidis, Hayman Lotfy, and Michalis Vlachos. 2026. Cost-Aware Human-LLM Collaboration for post-OCR Corrections in Swiss Historical Newspapers. In *Proceedings of the 2026 ACM Symposium on Document Engineering (DocEng '26), August 25–28, 2026, Fribourg, Switzerland*. ACM, Fribourg, Switzerland, 4 pages. https://doi.org/10.1145/3820755.3832797

1 Introduction

Historical newspaper archives are invaluable for humanities research, yet their OCR transcriptions often suffer from high Character-error rates (CER) caused by old typefaces, physical degradation, irregular spelling, and complex layouts [1, 11, 19]. Large Language Models (LLMs) can correct many of these errors contextually; however, applying them indiscriminately is costly, unnecessary for clean segments, and risks introducing "modernization" errors that alter historical spelling conventions (see Fig. 2).

To address these issues, we propose a three-tier collaboration framework that routes each raw OCR segment to one of: (1) No Correction, retaining the OCR text; (2) LLM Correction, applying automated post-editing; or (3) Human Correction, delegating highly degraded text for expert review. We evaluate our routing strategies on a historical Swiss newspaper corpus (1733–1945), predicting optimal delegation paths from raw OCR features alone.

Our main **contributions** are:

- A cost-aware routing framework coordinating OCR outputs, LLMs, and human reviewers, with a regression-guided routing strategy.
- A routing approach with a safeguard layer that detects when an LLM correction will deteriorate the OCR result and routes uncertain segments to human review under a small budget constraint (<5% of the corpus).
- An evaluation on a historical Swiss newspaper corpus (1733–1945) showing that our method achieves 2.55% CER with only <5% of the corpus reviewed by human experts, a 14% relative reduction over the All-LLM baseline (2.98% CER) and substantially outperforming confidence-based approaches.

2 Related Work

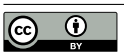
Post-OCR correction is transitioning from rule-based and statistical methods, such as noisy-channel models [12], to deep learning and Large Language Models [15]. Recent studies show that LLMs can improve historical newspaper transcriptions [23], though performance remains sensitive to prompt design, language, and script complexity [4, 22]. Complementary approaches include byte-level language models for monolingual spelling correction [3] and LLM-based correction pipelines for low-resource historical scripts such as Old Greek [6]. While traditional dictionary-based spell-checkers struggle with real-word errors (which account for up to 59.0% of OCR noise [10]), they can also trigger modernization traps by changing valid archaic spellings [9].

Confidence-Aware Detection and Hybrid Models. OCR engine confidence scores are frequently used to locate errors. Hemmer et al. [10] propose ConfBERT, which fuses token confidence scores with language representations for binary error detection. Unlike token-level error detection, our framework operates at the document-segment level to make three-class routing decisions based on surface, spelling, and metadata features.

Cost-Aware Routing. As LLM adoption grows, cost-aware routers like RouteLLM [13, 16] learn to route queries between strong and weak models. Similarly, bandit-based frameworks choose among multiple LLMs [18, 24]. Complementary to this, our work addresses the decision of whether a correction is needed, and *if* an LLM correction is adequate.

3 Methodology

We frame OCR post-correction as a three-tier *human-LLM collaboration* problem: rather than correcting every segment using either



DocEng '26, Fribourg, Switzerland

ACM ISBN 979-8-4007-2786-3/2026/08
https://doi.org/10.1145/3820755.3832797

the LLM or a human reviewer, we choose the most cost-effective path for each segment using only features available from the raw OCR output (see Fig 1). This section details our collaboration tiers, the cost model, the threshold parameter, and the learned routing classifiers.

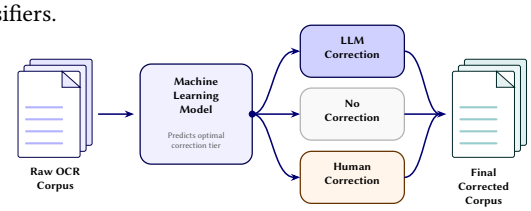


Figure 1: Overview of our three-tier, cost-aware human-LLM collaboration framework.

Three-Tier Collaboration Model. Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ be a corpus of segments. Each segment d_i has a raw OCR transcription d_i^{raw}, a potential LLM-corrected version d_i^{corr}, and a ground-truth human-corrected version d_i^{human} (assumed to have zero errors). To handle severely degraded content, we explicitly include OCR failures where the engine produced no output, setting $d_i^{\text{raw}} =$ "" with a default raw $CER(d_i^{\text{raw}}) = 1.0$. We define three correction paths based on the raw and corrected error rates:

- **No Correction (Class 0):** Bypassed if $CER(d_i^{\text{raw}}) = 0$ to avoid API cost and prevent modernization errors.
- **LLM Correction (Class 1):** Accepted if $CER(d_i^{\text{raw}}) > 0$ and the LLM brings the residual error under a quality threshold x ($CER(d_i^{\text{corr}}) \leq x$).
- **Human Correction (Class 2):** Segments routed for human review if $CER(d_i^{\text{raw}}) > 0$ and the LLM's corrected version remains poor ($CER(d_i^{\text{corr}}) > x$).

These labels are used to characterize the ideal correction path during evaluation. At deployment time, the true error rates are unavailable; the learned router approximates these decisions using only features extracted from the raw OCR segment.

Cost Model. Segments represent newspaper blocks averaging 220 characters (~35 words). Class 0 carries zero cost. Class 1 (LLM) carries an API cost. Class 2 (Human) carries an estimated time cost of $T_{\text{human}} = 4$ minutes per segment (15 segments/hour) to compare the OCR with the historical page image. For an hourly wage W, the human segment cost is $C_{\text{human}} = W/15$ (e.g., \$1.00 at \$15/hour wage, \$2.00 at \$30/hour wage [1]). Under current pricing, the LLM correction cost is approximately three orders of magnitude smaller than the human cost, making the routing decision almost entirely a question of human workload minimization.

OCR, LLMs, and Prompt Design. To evaluate correction generalizability, we run three OCR engines: PaddleOCR v3 [5], EasyOCR[2], and Tesseract v5 [21]. Post-correction is performed using six LLMs, spanning closed-source (Gemini 3 Flash, GPT-4o, Claude 3.5 Haiku)[2, 7, 17] and open-weight (Qwen 2.5-72B, Llama 3.3-70B, Gemma 4 31B) [8, 14, 20] models. For each model, we evaluate ten distinct prompt templates of increasing complexity: (i) *Basic* (Prompts 1–2) provide direct, zero-shot instructions focused on raw text recovery, spacing, and punctuation; (ii) *Intermediate* (Prompts 3–6) introduce historical rules to preserve archaic French orthography and resolve standard typographic confusions (e.g., long 's' f/s confusion); (iii) *Expert/Master* (Prompts 7–10) integrate in-context few-shot exemplars, structured Chain-of-Thought reasoning steps, and strict output formatting constraints to eliminate conversational preambles. Due to space limits, we report results using Tesseract and Gemini 3 Flash, which represent the optimal quality-cost configuration; results for other models and the full prompt ablation study are available in our code repository.

Regression-Guided Routing with Safeguard. Rather than casting routing as classification, we adopt a two-step approach:

Step 1: Priority ordering. A LassoCV regression model predicts the continuous improvement in CER ($\Delta\text{CER} = \text{CER}_{\text{raw}} - \text{CER}_{\text{corr}}$) from the raw-OCR feature vector $\phi(d_i)$. A positive ΔCER indicates that LLM correction reduces error. Segments are sorted by predicted ΔCER (most beneficial first) and corrected incrementally.

Step 2: Per-segment safeguard. As each segment is processed in the priority order, a safeguard classifier decides its correction path. We train a logistic regression model under 10-fold CV to estimate $P(\Delta\text{CER} \geq 0 \mid \phi(d_i))$ the probability that the LLM correction is safe (i.e., does not increase error). For each segment:

- If $P \geq 0.5$: the LLM correction is **accepted** (correction is likely safe).
- If $P < 0.5$ and the human budget B is not exhausted: the segment is **routed to human review** (perfect correction, CER = 0).
- If $P < 0.5$ and the budget is exhausted: the segment **falls back to LLM correction**.

We set $B < 5\%$ of the corpus. This design ensures that the safeguard can only improve upon the unguarded routing curve, since it replaces a subset of potentially harmful LLM corrections with perfect human corrections. All models are evaluated under 10-fold cross-validation on our segments.

4 Experimental Results

We structure the evaluation around two research questions:

- **RQ1:** How effectively does regression-guided routing prioritize LLM corrections, and can a safeguard with a small human budget improve quality?
- **RQ2:** Can post-LLM safeguard classifiers detect and intercept harmful corrections under a fixed operational budget?

All code, data, prompts, and per-model results are available in our public repository.[3]

Corpus and annotation. We use a corpus drawn from nine distinct periodicals published in the Canton of Vaud between 1733 and 1945, spanning over two centuries of French-language press, from 18th-century literary journals (Mercure Suisse) to 20th-century dailies (Tribune de Lausanne) and totaling 609 text segments across 45 digitized pages. This temporal and typographic diversity ensures that the evaluation covers a wide range of OCR degradation patterns, including early typefaces (see Fig. 2), 19th-century serif layouts, and modern multi-column newspaper formats. Three French-speaking

[1] Costs reported for freelance annotators: https://freelancer.ypai.ai/jobs/french/annotation, as of April 2026

[2] https://github.com/JaidedAI/EasyOCR

[3] https://github.com/Stergios-Konstantinidis/Cost-Aware-Human-LLM-Collaboration-for-post-OCR-Corrections

annotators produced fidelity-oriented transcriptions that preserve historical spelling and source anomalies. We report WER and CER, computed after a topological normalization step that unifies historical character variants and spacing.

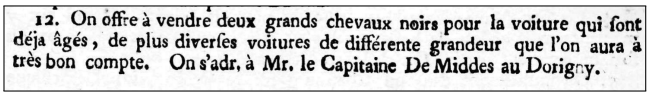

12. On offre à vendre deux grands chevaux noirs pour la voiture qui ſont déja âgés, de plus diverſes voitures de différente grandeur que l'on aura à très bon compte. On s'adr. à Mr. le Capitaine De Middes au Dorigny.

Figure 2: Example text segment from the *Feuille d'Avis de Lausanne* (1762). The long-*s* typeface causes OCR engines to read "diverfes" (*diverses*) and "font" (*sont*). "au Dorigny" would in modern grammar be over-corrected to "à Dorigny"

Baselines and protocol. We compare our regression-guided routing approach against three baselines: (1) Raw OCR (no further correction), (2) All-LLM Correction, and (3) Oracle (perfect routing by true ΔCER). We also compare against ConfBERT [10], a confidence-based error detection baseline. For regression-guided routing, we evaluate under 10-fold cross-validation over segments using the 54-dimensional surface, spelling, and metadata feature set. For the post-LLM safeguard experiment (RQ2), we evaluate on the full 609 segment corpus (where 12 OCR scans consistently fail and return empty strings). For each routing decision, we simulate the resulting text quality: segments routed to *No Correction* retain the raw OCR text, segments routed to *LLM Correction* use the actual Gemini 3 Flash corrected text, and segments routed to *Human Correction* are assumed to be perfectly corrected ($WER = 0, CER = 0$).

RQ1: Regression-Guided Routing Frontier. Figure 3 compares CER routing frontiers as segments are incrementally corrected. The Oracle (perfect routing) sorts by true ΔCER and achieves the steepest descent. Our LassoCV approach (faded red) closely tracks the Oracle, substantially outperforming the ConfBERT baseline [10] across the entire correction range. At 100% correction, the LLM reduces baseline CER from 6.32% to approximately 3.0%.

Adding the safeguard with a human-review budget capped at <5% of the corpus further improves quality. The safeguard curve (solid blue) lies at or below the unguarded routing curve at every operating point, confirming that routing flagged segments to human review strictly improves quality. With fewer than 5% of segments reviewed by an expert, the safeguard closes the gap toward the Oracle bound, demonstrating the cost-effectiveness of targeted human intervention on the most uncertain cases.

Analyzing the LassoCV coefficients (Figure 4) reveals which raw OCR features best predict improvement from LLM correction. Average OCR confidence (`avg_confidence`, $\beta = -0.040$) is the dominant predictor: segments with high confidence have little room for improvement, so the model correctly deprioritizes them. Among surface features, character frequency of 'y' ($\beta = +0.011$) and space ratio ($\beta = +0.004$) are the strongest positive predictors of ΔCER, indicating segments where the LLM yields the largest gains. Only 11 of 54 features receive non-zero coefficients, confirming that the Lasso's sparsity effectively selects a compact, interpretable routing signal.

RQ2: Safeguard Routing. Table 1 quantifies the routing performance at key correction budgets. At every operating point, our LassoCV approach outperforms the ConfBERT baseline. Adding the safeguard, a logistic regression classifier trained on the same

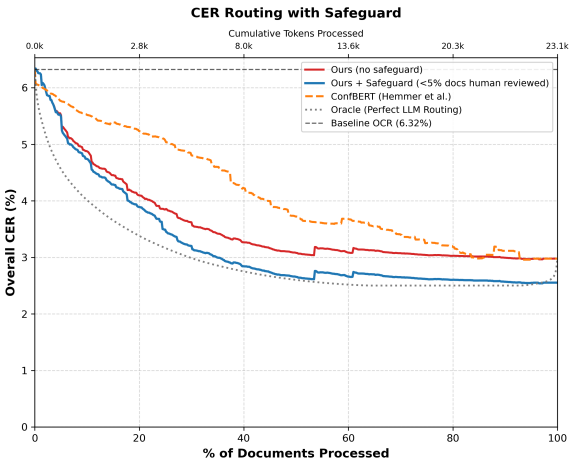


Figure 3: CER routing frontier comparing our LassoCV approach (with and without safeguard), ConfBERT [10], and Oracle (perfect routing). The safeguard routes <5% of segments to human review.

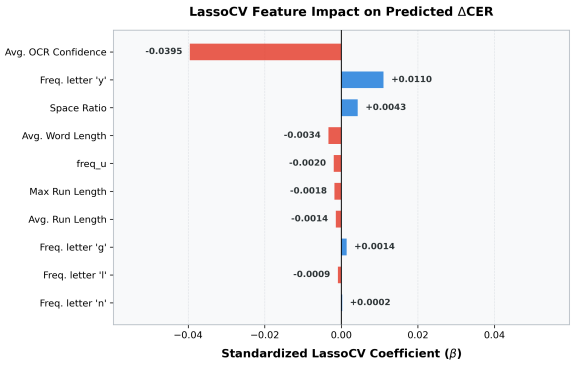


Figure 4: Standardized LassoCV coefficients for predicting ΔCER. Positive coefficients (blue) indicate features where the LLM correction yields larger gains; negative coefficients (red) indicate features negatively correlated with CER improvement.

raw OCR features to predict $P(\Delta\text{CER} \geq 0)$, further improves quality by routing the most uncertain corrections to human review (<5% of the corpus, 29 segments). At 50% correction, the safeguard achieves **2.66% CER** vs. 3.08% without safeguard and 3.72% for ConfBERT. At full correction, the safeguard yields **2.55% CER**, a 14% relative improvement over the All-LLM baseline (2.98%), with fewer than 5% of segments requiring expert review. These results confirm that targeted human intervention on the most uncertain cases strictly improves quality over unguarded LLM correction.

5 Discussion

The routing decisions in our framework are ultimately governed by the cost ratio between human and LLM correction. At current pricing, this ratio is striking: correcting all 609 segments with Gemini 3 Flash costs approximately $0.24 in API fees, whereas hiring a human annotator at $15/hour to review the same corpus at 15

Table 1: Overall CER (%) [95% bootstrap confidence intervals] at key correction budgets across routing strategies. Ours + Safeguard routes <5% of segments to human review. Oracle represents LLM correction only. Significance comparison (Wilcoxon signed-rank test): *,**,*** indicate $p < 0.05, 0.01, 0.001$ vs. ConfBERT; † indicates $p < 0.001$ vs. Ours.

% Corrected	ConfBERT	Ours	Ours + Safeguard	Oracle
0%	6.32 [5.38, 7.37]	6.32 [5.38, 7.37]	6.32 [5.38, 7.37]	6.32 [5.38, 7.37]
25%	5.07 [4.23, 6.02]	3.85* [3.20, 4.59]	**3.49****† [2.91, 4.16]	3.17 [2.61, 3.82]
50%	3.72 [3.03, 4.51]	3.08 [2.47, 3.80]	**2.66*****† [2.13, 3.30]	2.60 [2.05, 3.26]
75%	3.25 [2.58, 4.03]	3.04 [2.36, 3.83]	**2.62*****† [2.01, 3.34]	2.50 [1.95, 3.15]
100%	2.98 [2.29, 3.77]	2.98 [2.29, 3.77]	**2.55*****† [1.94, 3.27]	2.98 [2.29, 3.77]

segments per hour would cost roughly $600 and require nearly 40 hours of expert time. The LLM correction cost is therefore approximately three orders of magnitude smaller than the human alternative, making the routing decision almost entirely a question of *when* human intervention adds enough quality to justify the expense.

Our results show that the answer is surprisingly targeted. The safeguard routes only 29 segments (<5% of the corpus) to human review, yet this small allocation reduces CER from 2.98% to 2.55%. Each human-reviewed segment contributes disproportionately to overall quality because it replaces a case where the LLM would have introduced errors.

As better LLM corrections become effectively free, the routing decision grows more complex: Should a segment go to a cost-effective model, a quality-optimized one, or a human expert? In this multi-tier setting, the safeguard's ability to predict $P(\Delta\text{CER} \geq 0)$ becomes the critical component, since it directly determines how many segments need expert review.

Conversely, if human correction costs decrease (through crowdsourcing, semi-automated review interfaces, or improved OCR that reduces the number of degraded segments) the optimal human budget B could increase, and the framework would naturally route more segments to human review. The operating-point flexibility of our approach accommodates this: practitioners can adjust B as a single parameter to reflect their cost structure, without retraining the routing model. This adaptability makes the framework robust to the rapid shifts in both LLM pricing and human labor costs that characterize the current landscape.

Acknowledgments

Supported by Swiss National Science Foundation, Grant No. 237991.

References

[1] Beatrice Alex, Claire Grover, Ewan Klein, and Richard Tobin. 2012. Digitised Historical Text: Does it have to be mediOCRe?. In *Proceedings of KONVENS 2012.*

[2] Anthropic. 2024. *The Claude 3 Model Family: Opus, Sonnet, Haiku.* Technical Report. Anthropic. https://assets.anthropic.com/m/1cd9d098ac3e6467/original/Claude-3-Model-Card-October-Addendum.pdf

[3] Sávio Santos de Araújo, Byron Leite Dantas Bezerra, and Arthur Flor de Sousa Neto. 2025. A Proposal of Post-OCR Spelling Correction Using Monolingual Byte-level Language Models. In *Proceedings of the 25th ACM Symposium on Document Engineering (DocEng '25).* ACM.

[4] Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study. In *Proceedings of Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage.* doi:10.18653/v1/2024.latechclfl-1.14

[5] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR 3.0 Technical Report. doi:10.48550/arXiv.2507.05595

[6] Andreas Evangelatos, Konstantinos Palaiologos, Basilis Gatos, Panagiotis Kaddas, Aikaterini Christopoulou, and Vassilis Katsouros. 2025. Old Greek OCR Result Correction Using LLMs. In *Proceedings of the 25th ACM Symposium on Document Engineering (DocEng '25).* ACM. doi:10.1145/3704268.3748676

[7] Google DeepMind. 2025. *Gemini 3 Flash Model Card.* Technical Report. Google. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf

[8] Google DeepMind. 2026. *Gemma 4 Model Card.* Technical Report. Google. https://blog.google/technology/developers/gemma-4/

[9] Elise Hammarström. 2025. *Investigation of OCR Model Performance and Post-OCR Correction Strategies: A Comparative Analysis.* Master's thesis. Uppsala University, Faculty of Science and Technology. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-565512 Available at DiVA, id: diva2:1991132.

[10] Arthur Hemmer, Mickaël Coustaty, Nicola Bartolo, and Jean-Marc Ogier. 2024. Confidence-Aware Document OCR Error Detection. In *Document Analysis Systems: 16th IAPR International Workshop.* doi:10.1007/978-3-031-70442-0_13

[11] Rose Holley. 2009. How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine* 15, 3/4 (2009). doi:10.1045/march2009-holley

[12] Ido Kissos and Nachum Dershowitz. 2016. OCR Error Correction Using Character Correction and Feature-Based Word Classification. In *12th IAPR Workshop on Document Analysis Systems.* doi:10.1109/DAS.2016.44

[13] Zheng Li, Xuyun Zhang, Sheng Lu, Hua Deng, Hao Tian, and Wanchun Dou. 2025. A Cost-Aware Approach for Collaborating Large Language Models and Small Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management.* doi:10.1145/3746252.3760990

[14] Meta AI. 2024. *Llama 3.3 Model Card.* Technical Report. Meta. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

[15] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of Post-OCR Processing Approaches. *Comput. Surveys* 54, 6, Article 124 (2021). doi:10.1145/3453476

[16] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, Mohammed Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to Route LLMs from Preference Data. In *International Conference on Learning Representations.* https://proceedings.iclr.cc/paper_files/paper/2025/file/5503a7c69d48a2f86fc00b3dc09de686-Paper-Conference.pdf

[17] OpenAI. 2024. GPT-4o System Card. https://arxiv.org/abs/2410.21276

[18] Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. 2025. Adaptive LLM Routing under Budget Constraints. *arXiv preprint arXiv:2508.21141* (2025).

[19] Bhawna Piryani, Jamshid Mozafari, Abdelrahman Abdallah, Antoine Doucet, and Adam Jatowt. 2025. Evaluating Robustness of LLMs in Question Answering on Multilingual Noisy OCR Data. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management.* doi:10.1145/3746252.3761295

[20] Qwen Team. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024). https://arxiv.org/abs/2412.15115

[21] R. Smith. 2007. An Overview of the Tesseract OCR Engine. In *International Conference on Document Analysis and Recognition*, Vol. 2. doi:10.1109/ICDAR.2007.4376991

[22] Muhammad Abdullah Sohail, Salaar Masood, and Hamza Iqbal. 2024. Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts. https://arxiv.org/abs/2412.16119

[23] Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages.* https://aclanthology.org/2024.lt4hala-1.14/

[24] Wang Wei, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A. Rossi, and Hoda Eldardiry. 2025. Learning to Route LLMs from Bandit Feedback: One Policy, Many Trade-offs. *arXiv preprint arXiv:2510.07429* (2025).