Digitizing Old Ukrainian Texts: A Prompt-Based OCR Pipeline and Evaluation Dataset

Dmytro Chaplynskyi
I. Krypiakevych Institute
of Ukrainian Studies, lang-uk
chaplinsky.dmitry@gmail.com

Hanna Dydyk-Meush
I. Krypiakevych Institute
of Ukrainian Studies of NAS of Ukraine
hanna.dydykmeush@gmail.com

Abstract

We present a methodology and an open dataset for OCR of handwritten index cards containing a scholarly transcription of an early 17th-century Ukrainian polemical text, *Perestoroha* by Iov Boretskyi (Lviv, 1605–1606). The 430 cards, produced by 20th-century researchers, preserve the text in Old Ukrainian orthography with archaic diacritics, titlos, superscript letters, and ligatures that make automated recognition non-trivial. We develop a prompt-based OCR pipeline driven by a custom instruction set designed iteratively from the source material's orthographic conventions. The pipeline is evaluated against human-proofread ground truth in proprietary and open-source configurations using identical instructions and evaluation data. The proprietary configuration with extended thinking at maximum budget (Claude Opus 4.7, xhigh) achieves a Character Error Rate of **2.5%**; an Opus 4.6 baseline at the default 2,048-token thinking budget—used for the first batch of the released dataset—reaches **4.2%**; and two open-source Qwen3.6 variants running locally on consumer hardware reach **14.6%** (dense 27B) and **14.8%** (35B-A3B MoE). We release the fully digitized text aligned at line level to 300 DPI scanned images, as both a scholarly digital resource and training data for future OCR systems targeting Old Slavic manuscripts.[1]

1 Introduction

A significant body of Old Ukrainian linguistic material remains accessible only in handwritten form—either as original manuscripts or as scholarly transcriptions produced by researchers over the past century. Digitizing these materials is a prerequisite for computational analysis, full-text search, lexicography, and long-term preservation, yet the archaic orthographic conventions they employ render standard OCR tools ineffective.

We address this problem for a specific artifact: a set of 430 handwritten index cards constituting a scholarly transcription of *Perestoroha zelo potrebnaia na potomnye chasy pravoslavnym khrystiianom*, a polemical text by Iov Boretskyi written in Lviv in 1605–1606. The cards were produced by 20th-century researchers who carefully reproduced the original text, preserving its Old Ukrainian orthography, titlo abbreviation marks, superscript letters, and other diacritical features.

Rather than fine-tuning a dedicated handwritten text recognition model—which would require substantial annotated training data—we adopt a **prompt engineering** approach: we develop a detailed instruction set that guides a multimodal large language model to transcribe each card image into structured text while preserving all orthographic features. The instruction set was developed iteratively against Claude Opus 4.6 (Anthropic, 2025) through analysis of the source material's character inventory, abbreviation conventions, and diacritic usage patterns; we then re-evaluate the same instruction set, without modification, on Claude Opus 4.6 (production transcription baseline), Claude Opus 4.7 with extended thinking, and on two Qwen (Yang et al., 2025) variants as open-source replications.

Our contributions are: (1) a practical, reproducible prompt-based OCR pipeline for handwritten scholarly transcriptions of Old Ukrainian texts, evaluated against proprietary and fully open-source LLMs; and (2) an open dataset of 430 card transcriptions aligned to high-resolution scans.

2 Related Work

Historical HTR systems. Handwritten text recognition for historical documents has advanced substantially, driven by deep learning and large-

[1] Code: https://github.com/lang-uk/slavon_ocr/ Dataset: https://huggingface.co/datasets/lang-uk/perestoroha-ocr

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 58–66

scale digitization. Transkribus (Muehlberger et al., 2019) is the most widely adopted platform, achieving CERs below 5% across hundreds of public models. Kraken and eScriptorium (Kiessling, 2019; Kiessling et al., 2019) offer a fully open-source alternative designed for non-Latin and historical scripts. TrOCR (Li et al., 2023) introduced a fully Transformer-based architecture, achieving state-of-the-art results on handwriting benchmarks. CHURRO (Semnani et al., 2025), a 3B-parameter VLM fine-tuned on 155 historical corpora spanning 46 language clusters and 14 scripts, represents the current state of the art in VLM-based historical text recognition.

Old Slavic HTR. Work on Old Slavic manuscripts remains sparse compared to Western European languages. Neural HTR for Church Slavonic was pioneered via Transkribus, achieving CERs of 3–5% and identifying key error sources—superscript letters, titlo abbreviations, and word separation—that directly correspond to error categories in our pipeline (Rabus, 2019). The first generic HTR model for Ukrainian handwriting (CER 4.2%) was trained on 19th–20th century manuscripts (Tikhonov and Rabus, 2024); no existing model covers the 17th-century orthography targeted here. Recent work on HTR postprocessing of pre-modern Slavic texts identifies the same challenges with superscript letters and titlos that we encounter (Lendvai et al., 2024). Critically, all existing approaches rely on Transkribus or Kraken models requiring annotated training data.

Multimodal LLMs for OCR. Recent work shows that multimodal LLMs can match or exceed specialized HTR systems. On 18th–19th century English documents, GPT-4o, Claude, and Gemini achieve CERs of 5.7–7% (Humphries et al., 2025), and systematic comparisons find that LLMs significantly outperform all conventional methods on historical handwritten records (Kim et al., 2025). A multilingual HTR benchmark shows Claude 3.5 Sonnet outperforms other models in zero-shot settings but with weaker non-English performance (Crosilla et al., 2025). For Slavic text specifically, an evaluation of 12 multimodal LLMs on 18th-century Cyrillic finds that models exhibit "over-historicization"—inserting archaic characters from incorrect periods (Levchenko, 2025).

Prompt engineering for documents. Layout-aware prompting improves document understanding by 263% in zero-shot settings (Wang et al., 2023). Comparisons of prompting strategies for historical handwriting find that detailed document descriptions with few-shot examples yield optimal results (Kim et al., 2025). Structured prompt engineering with JSON output schemas enables effective extraction from specialized text without fine-tuning (Dagdelen et al., 2024). In low-resource settings, where annotated training data for HTR is unavailable, recent benchmarks of LLM-based OCR on under-served scripts (Sohail et al., 2024) report that careful prompt design closes a substantial part of the gap to specialized systems—a finding that motivates the instruction-set engineering approach we adopt for 17th-century Ukrainian.

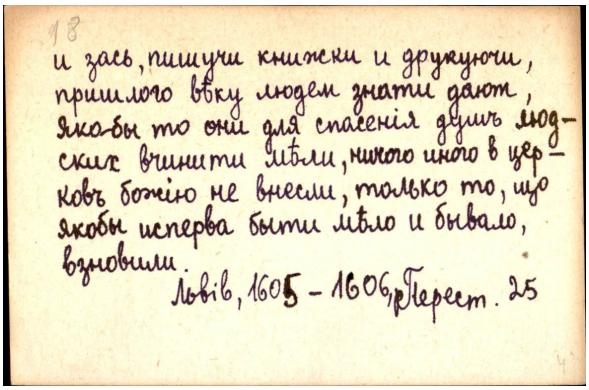

Figure 1: Example index card from the *Perestoroha* collection. The handwriting preserves 17th-century orthographic conventions. The bottom line shows the source reference (Lviv, 1605–1606, Perest. 25).

3 The Perestoroha Image Dataset

The source material consists of 430 index cards, scanned at 300 DPI, all belonging to a single scholarly artifact. Each card reproduces a fragment of *Perestoroha* with source attribution noting library provenance, approximate date, and folio references. The cards were handwritten by 20th-century researchers, making the handwriting itself relatively legible; the difficulty lies in the orthographic system being reproduced.

The text employs Old Ukrainian orthography of the 17th century. Key features include: ѣ (yat), ѡ (omega as ot-ligature), ї (yi with two dots), ѫ (big yus); **titlo** marks over abbreviated sacra nomina; superscript letters indicating abbreviation expansions; acute stress marks on vowels; and the coexistence of є and е, ꙋ and у—visually similar pairs requiring per-instance discrimination rather than any default substitution rule.

4 OCR Pipeline

4.1 Overview

The pipeline consists of five stages: (1) splitting scanned PDF files into individual card images; (2) an EXIF-orientation preprocessing pass that bakes any rotation flag into the pixel buffer (§4.5); (3) instruction-driven multimodal LLM transcription of each image into structured JSON; (4) import into a web-based proofreading editor for expert review; and (5) export of corrected transcriptions into a browsable dataset. Each card is transcribed independently—no context carries over between cards—ensuring reproducibility and enabling parallel processing. The code is available at `https://github.com/lang-uk/slavon_ocr/`.

4.2 The Instruction Set

The core of the pipeline is a custom instruction set for Claude Opus 4.6 (Anthropic, 2025), developed iteratively over multiple rounds of analysis and error correction. Its design drew on several sources:

- **Three ground-truth cards** with matched manual transcriptions, which revealed the scholar's encoding conventions: parentheses for expanded abbreviations, specific Unicode codepoints for diacritical marks, and character-level encoding decisions.
- **A 12,000-word reference corpus** of the same text (already digitized[2]), which yielded a full character inventory, uncovering characters initially missed—Latin *s* for zelo, the positional distribution of ꙋ vs у, and a complete inventory of sacra nomina with titlos.
- **Lecture materials** on Old Ukrainian phonology and orthography, which contributed precise descriptions of rarely encountered marks: paieryk, kamora, and prydykh.

The instruction set specifies a detailed character encoding table with Unicode codepoints, rules for diacritic handling, the convention for representing abbreviation expansions, and a structured JSON output schema. Key components are provided in Appendix A.

[2]The digitized portion of Кройніка used here as a reference corpus results from two phases of philological work. In 1981, the linguists Valentyna Cherniak and Oleksandra Zakharkiv selectively transcribed the monument for the Card Index of the *Dictionary of the Ukrainian Language of the 15th – first half of the 17th centuries*. In 2013–2014, Hanna Dydyk-Meush and Oksana Shpyt worked with the original of Кройніки and transcribed an extended portion in full; that latter transcription is what we use here.

4.3 Anti-Hallucination Measures

Early experiments revealed a critical failure mode: the model substitutes familiar words for unfamiliar archaic ones when letter shapes are ambiguous. For example, Спудєй was misread as Судєй, and испервa as неперва—plausible modern forms replacing rare historical ones.

We adopted a three-pronged mitigation strategy: (1) explicit "read letter by letter, not word by word" instructions; (2) a verification step listing concrete examples of documented error types (є/е confusion, ꙋ/у substitution, spurious trailing ъ); and (3) a "no defaults" policy—both members of each visually similar pair (є/е, ꙋ/у) are flagged as requiring per-instance verification against the card image.

4.4 Thinking Spiral Mitigation

With extended thinking (chain-of-thought reasoning) enabled at large reasoning budgets on Claude Opus 4.6, we observed a failure mode in which the model entered an unbounded deliberation loop on a card with many ambiguous characters, consuming its entire output token budget without producing any transcription. This occurred at both 4K and 32K thinking budgets. We did *not* observe the issue on *Perestoroha*—all 152 cards in the proofread corpus completed via the thinking-on path on both the Opus 4.6 production run (default 2,048-token thinking budget) and the Opus 4.7 evaluation (xhigh budget)—but it surfaced on a different handwritten source we work with, and colleagues working with other handwritten Slavic sources have reported the same pattern privately, suggesting it is not corpus-specific.

The root cause is that extended thinking expands to fill whatever budget is available, and the instruction set's emphasis on precision amplifies this tendency. Our pipeline therefore applies a budget-aware retry strategy: each card is first transcribed with extended thinking enabled at the configured budget; if the call exceeds the budget without producing a transcription, the card is retried with extended thinking disabled. On *Perestoroha* this fallback is never invoked, but it degrades gracefully on the small number of cards prone to deliberation loops in other corpora. A prompt-level mitigation—a "deliberation discipline" section instructing the model to make one pass, verify once, and commit—proved helpful but insufficient on its own.

4.5 EXIF Orientation Preprocessing

A non-trivial fraction of the scans—roughly 25% (over 100 files)—carried EXIF rotation flags whose target orientation differed from the underlying pixel buffer. Multimodal models honor these flags at decode time, so the model was shown the image rotated 90° or 180° from the writer's intended orientation. On the worst-affected cards this was catastrophic: a single card hit a CER of approximately 137%, with the model producing essentially gibberish in an attempt to read inverted handwriting. We added a one-time prepass that reads the EXIF tag, applies the implied rotation to the pixel data itself, and strips the metadata; both Claude and Qwen runs consume the resulting auto-oriented image directory. After the fix, the previously affected cards dropped to in-distribution CERs in the 3–5% range. In an ablation against the proofread evaluation corpus on the Opus 4.6 production configuration, removing the prepass raised aggregate CER from 4.16% to approximately 5.3% (+1.1 pp) and WER by several points, because the long tail of mis-rotated cards contributes disproportionately to both metrics: each misread compounds with re-segmentation noise across subsequent lines.

4.6 Proofreading Editor

We built a custom web-based proofreading editor (Flask + SQLite) using Claude Code. The editor displays the scanned card image alongside the model's transcription in a side-by-side view (Figure 2), with a toolbar for inserting Old Ukrainian diacritics and archaic characters, enabling a domain expert to correct errors character by character. Corrected transcriptions are exported as parallel JSON files, preserving the original OCR output for comparison. Although each card's initial draft was produced by Claude, every character in the resulting ground truth was verified against the source image by the annotator; the proofread reference is therefore independent of the model's lexical choices, even where the draft seeded the typing.

Proofreading was performed by a co-author with domain expertise in historical Ukrainian linguistics. The annotator's familiarity with the source material's orthographic system was sufficient to resolve ambiguities without formal adjudication guidelines.

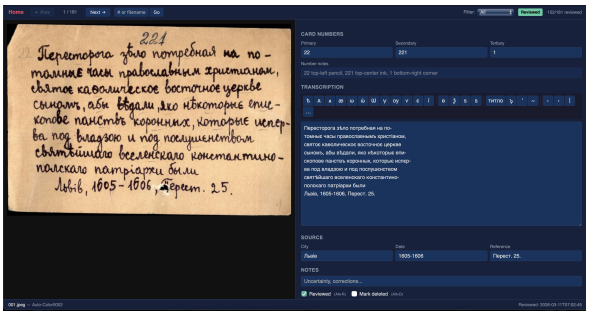

Figure 2: The proofreading editor showing a card image (left) alongside the model's transcription (right), with fields for card numbers, source metadata, and a diacritics toolbar.

5 Evaluation

5.1 Experimental Setup

We evaluate the pipeline in three configurations using identical instructions and evaluation data:

- **Claude Opus 4.7 with extended thinking** (Anthropic, 2025)—proprietary, API access, maximum reasoning budget (`xhigh`)—our headline configuration, applied to the proofread evaluation set.
- **Claude Opus 4.6** (Anthropic, 2025)—proprietary, API access, extended thinking enabled at the default 2,048-token budget—the production configuration used for the first batch of cards in the released dataset, reported here as a same-prompt baseline against which to measure the impact of the model upgrade.
- **Two open-source Qwen3.6 variants** (Yang et al., 2025), run locally on a single NVIDIA RTX 4090 GPU via `llama.cpp`: a dense **Qwen3.6-27B** model and a **Qwen3.6-35B-A3B** mixture-of-experts model with 3B active parameters per token. Both use 4-bit Q4_K_M GGUF quantizations from the Unsloth distribution.[3] The same instruction set was applied to both without modification.

5.2 Metrics

We report Character Error Rate (CER) and Word Error Rate (WER). CER is the primary metric, as word boundaries in this orthographic system are not always unambiguous, and many errors involve single-character substitutions within diacritical marks. Tokenization for WER uses the spaCy

[3] https://huggingface.co/unsloth/Qwen3.6-27B-GGUF and https://huggingface.co/unsloth/Qwen3.6-35B-A3B-GGUF.

Configuration	CER	WER
Claude Opus 4.7 (xhigh thinking)	2.51%	12.16%
Claude Opus 4.6 (default 2K thinking)	4.16%	18.10%
Qwen3.6-27B (dense, local)	14.56%	39.44%
Qwen3.6-35B-A3B (MoE, local)	14.84%	40.34%

Table 1: OCR accuracy on the proofread 152-card evaluation set, identical instruction set across configurations. WER tokenisation uses spaCy `uk_core_news_sm`; whitespace and punctuation tokens are dropped.

`uk_core_news_sm` pipeline; whitespace and punctuation tokens are dropped before the edit-distance computation, and hyphenated line-break splits are not rejoined.

5.3 Results

Results are shown in Table 1. Claude Opus 4.7 with extended thinking achieves 2.51% CER (542 character errors over 21,635 reference characters; 432 word errors over 3,552 word tokens), comparable to Transkribus-based results on Old Slavic texts (Rabus, 2019; Tikhonov and Rabus, 2024). Holding the instruction set fixed, switching from Opus 4.6 at the default 2K thinking budget (our production configuration) to Opus 4.7 with xhigh thinking reduces total character-level edits from 901 to 542—a 40% relative drop—and lifts the share of perfect cards (CER = 0) from 11.2% to **12.5%**. The two Qwen variants land within a hair of each other at 14.56% / 14.84% CER—roughly 6× higher than Opus 4.7—with the smaller dense 27B narrowly outperforming the larger MoE variant on every aggregate metric. A small tail of failure-mode cards drives most of the gap to Claude (§5.4).

An additional reference point: GPT-5.4 via Codex. As a sanity check on whether the proprietary advantage was Claude-specific, we ran the same instruction set through OpenAI's Codex CLI driving GPT-5.4 at the highest reasoning-effort setting on the 30-card test split. It reached a CER of 5.13% and WER of 24.08%, against 2.88% / 13.96% for Opus 4.7 (xhigh thinking) and 3.39% / 15.34% for Opus 4.6 on the same split, while taking roughly an order of magnitude longer per card in wallclock time than our Opus 4.6 production harness (we did not directly time-match it against the Opus 4.7 evaluation harness, which uses different parallelism). Although competitive with Opus 4.6 in absolute accuracy terms, GPT-5.4 offered no accuracy advantage on this task at substantially higher inference latency, so we did not pursue a full-corpus evaluation of this configuration.

5.4 Error Analysis

Character-level edit operations for Claude Opus 4.7 (xhigh thinking) on the 152-card proofread corpus comprise 542 total edits over 21,635 reference characters (376 substitutions, 63 insertions, 103 deletions). For comparison, the same instruction set on Opus 4.6 at the default 2K thinking budget produces 901 edits (550 / 193 / 158)—a 40% relative reduction at corpus scale, distributed unevenly across error classes. The most frequent residual error categories on Opus 4.7 are:

1. є/е **confusion** (U+0454 ↔ U+0435)—159 substitutions in both directions (є→е: 122; е→є: 37), accounting for 42% of all substitutions on Opus 4.7. The absolute count barely moves between Opus 4.6 (164) and 4.7 (159), while every other error class shrinks substantially—a strong indicator that this confusion reflects genuine visual ambiguity in the scholar's handwriting rather than a model failure mode that further prompt engineering will reduce.
2. **Spurious** ъ **insertion**—20 inserted hard signs, still the single largest non-whitespace insertion category, but down from 125 on Opus 4.6 (∼6× reduction). The model continues to over-apply the convention of word-final ъ from Church Slavonic templates, but only on a small residual set of cases.
3. ү/у **normalization** (U+04AF → U+0443)—27 cases where the rare straight-у variant is flattened to the dominant у, consistent with bias toward the more frequent letter form.
4. і/и **substitution** (U+0456 → U+0438)—13 cases where dotted і is read as plain и; plain и is roughly 4× more common in *Perestoroha* than і.
5. ѣ **(yat) misreadings**—10 substitutions spread across seven distinct target characters, reflecting the visual similarity of yat to several other letters when the writer's stroke is rushed.

Qwen-specific patterns. Both Qwen variants exhibit the same є/е confusion as Claude at substantially higher rates—173 substitutions on the 27B (є↔е sum), versus 159 on Opus 4.7—and a much stronger pull toward word-final ъ: 165 spurious insertions on the 27B and 161 on the 35B-A3B,

against just 20 on Opus 4.7 (∼8× more). They also show a modern-Ukrainian-bias pattern almost absent from the Claude error profile, including о→а substitutions (32 / 35 cases for 27B / 35B-A3B) and ы→и (21 / 39 cases for 27B / 35B-A3B). Severe whole-word hallucinations—the model generating plausible-looking Old Ukrainian text rather than transcribing what is on the card—are concentrated on a small tail of cards: per-card CER exceeds 50% on only 4/152 cards for the 27B and 3/152 for the 35B-A3B (roughly 2–3% of the evaluation). When those outlier cards are excluded, both Qwen variants produce CERs of approximately **8.8%**, suggesting that the aggregate gap to Claude is driven less by systematic recognition error than by a small number of failure-mode cards. The 35B-A3B's MoE architecture (3B active per token) produces marginally more deletions and fewer fabrications than the dense 27B, but the two are otherwise functionally equivalent on aggregate metrics. This comparison may not be entirely fair: the instruction set was developed for Claude, and both Qwen configurations used aggressively quantized GGUF builds. Prompt tuning specifically for the open-source models, or using less quantized variants, could plausibly close the residual gap further.

6 Dataset

We release the fully digitized text of *Perestoroha* as transcribed from 430 index cards.[4] The first batch of approximately 150 cards was transcribed with Opus 4.6 using an earlier version of the instruction set; the remainder was transcribed with Opus 4.7 (xhigh thinking) using the final instruction set described in §A. The dataset consists of JSON files paired with 300 DPI card images, aligned at line level. Each JSON record contains card numbers (primary, secondary, tertiary where present), an array of transcribed lines preserving exact line breaks for image-text alignment, source metadata (provenance, date, folio), and free-text notes flagging uncertainty.

The dataset serves two purposes: (1) as a **historical linguistic resource**—a searchable, citable digital text of a significant early 17th-century Ukrainian source that is otherwise difficult to access; and (2) as **OCR training and evaluation data** for future systems targeting Old Slavic Cyrillic handwriting.

[4]Dataset: https://huggingface.co/datasets/lang-uk/perestoroha-ocr

7 Discussion

Prompt engineering as a lightweight alternative to fine-tuning. The instruction set—approximately 3,000 words of encoding rules and error mitigation—required no training data beyond a few reference cards and a character inventory corpus. This makes the approach accessible to digital humanities practitioners who lack the resources for model fine-tuning.

Open-source viability. The two Qwen variants achieve CERs of 14.56% (dense 27B) and 14.84% (35B-A3B MoE) using the same instruction set without modification, with the smaller dense 27B narrowly outperforming the larger MoE variant on every aggregate metric—a useful finding for institutions that cannot use proprietary APIs and prefer a smaller, simpler deployment. While the aggregate numbers are still too high for unsupervised digitization, when the small tail of severely-hallucinated cards is excluded (∼2–3% of the evaluation) both variants produce CERs of approximately 8.8%, suggesting that the gap to Claude is driven primarily by failure-mode cards rather than by systematic recognition errors. This makes the open-source configuration a viable first pass for reducing manual transcription effort, and the released dataset is now large enough to support targeted fine-tuning aimed at suppressing the hallucination tail.

Broader implications. Ukrainian archives hold substantial collections of handwritten linguistic material from the 16th–18th centuries. The prompt-based approach demonstrated here can be adapted to other artifacts by developing source-specific instruction sets, potentially enabling large-scale digitization without the overhead of training dedicated models for each orthographic convention.

8 Conclusion

We presented a prompt-based OCR pipeline for digitizing handwritten scholarly index cards containing a 17th-century Ukrainian text. The pipeline, driven by a carefully engineered instruction set, achieves a CER of 2.5% with Claude Opus 4.7 (extended thinking), 4.2% with the Opus 4.6 production configuration, and 14.6–14.8% with two open-source Qwen3.6 variants run locally—improving to ∼8.8% once a small tail of failure-mode cards is excluded. We release the fully digitized text of *Perestoroha*—430 cards aligned to scanned

images—as an open resource for Ukrainian historical linguistics and OCR research.

Future work includes extending the pipeline to additional artifacts with different orthographic conventions (experiments are underway and show promising results), fine-tuning an open-source model on the released dataset, and building a searchable lexicographic resource from the digitized text.

Limitations

The pipeline was developed and evaluated on a single artifact with a specific orthographic system. Generalization to other Old Ukrainian texts—particularly those from different centuries or with different transcription conventions—requires partial re-engineering of the instruction set. The ground truth was produced by a single domain expert, and inter-annotator agreement has not been measured. The instruction set itself was iteratively refined—using Claude Code as a development assistant—against a 32-card sample drawn from the dev split during development on Opus 4.6, in order to reduce CER and WER on recurring failure modes; the 30-card test split was held out throughout. The full-corpus Opus 4.6 numbers therefore mix held-out and in-distribution performance, and readers wanting a pure held-out estimate should consult the test-split numbers in §5.3. The Opus 4.7 result is computed by running the same unchanged prompt on a newer model that the prompt was *not* tuned against, so the 4.6 → 4.7 delta partly reflects this looser coupling. The dataset is split deterministically into a 122-card development split and a 30-card test split (seed 20250101); the per-split numbers used for the GPT-5.4 comparison are reported separately to support cleaner held-out comparisons in future work. The proprietary configuration (Claude Opus) incurs API costs that may be prohibitive for large-scale digitization; the open-source alternative (Qwen) offers a cost-free option but at substantially lower accuracy, with a small tail of whole-word hallucinations and stronger biases toward Church Slavonic and modern-Ukrainian normalization that limit its utility without expert proofreading.

Ethical Considerations

This work involves digitization of historical scholarly materials in the public domain. No personal or sensitive data is processed. We used AI-based writing assistance tools (Claude) in preparing this manuscript, as well as Claude Code for developing the OCR pipeline and the proofreading interface. The open-source release of the dataset and instruction set is intended to enable reproducibility and further research.

Acknowledgments

We thank the staff of the I. Krypiakevych Institute of Ukrainian Studies of the National Academy of Sciences of Ukraine for access to the *Perestoroha* card collection and for support of this work. We thank Oksana Shpyt for her work on the philological transcription of Кройніка that produced our digitized reference corpus, Yurii Paniv for assistance with the Qwen experiments, and Mariana Romanyshyn for reviewing early drafts of this manuscript.

References

Anthropic. 2025. Claude Opus 4 system card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-opus-4-system-card.pdf.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition. *Journal of Documentation*, 81(7):334–354.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods*, 58(3):175–193.

Benjamin Kiessling. 2019. Kraken — a universal text recognizer for the humanities. In *Digital Humanities 2019 Book of Abstracts*, Utrecht.

Benjamin Kiessling, Robin Tissot, Peter A. Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An open source platform for historical document analysis. In *2019 ICDAR Workshops*, pages 19–24.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records. *arXiv preprint arXiv:2501.11623*.

Piroska Lendvai, Maarten van Gompel, Anna Jouravel, Elena Renje, Uwe Reichel, Achim Rabus, and Eckhart Arnold. 2024. A workflow for HTR-postprocessing, labeling and classifying diachronic and regional variation in pre-modern Slavic texts. In *Proceedings of LREC-COLING 2024*, pages 2039–2048. ELRA and ICCL.

Maria Levchenko. 2025. Evaluating LLMs for historical document OCR: A methodological framework for digital humanities. In *Proceedings of LM4DH 2025*, pages 75–85.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of AAAI 2023*, volume 37, pages 13094–13102.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, and 35 others. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Achim Rabus. 2019. Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus. *Scripta & e-Scripta*, 19:9–32.

Sina J. Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica S. Lam. 2025. CHURRO: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of EMNLP 2025*.

Muhammad Atif Sohail, Sarfaraz Masood, and Hammad Iqbal. 2024. Deciphering the underserved: Benchmarking LLM OCR for low-resource scripts. *arXiv preprint arXiv:2412.16119*.

Alexey Tikhonov and Achim Rabus. 2024. Handwritten text recognition of Ukrainian manuscripts in the 21st century: Possibilities, challenges, and the future of the first generic AI-based model. *Kyiv-Mohyla Humanities Journal*, 11:226–247.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023. Layout and task aware instruction prompt for zero-shot document image question answering. *arXiv preprint arXiv:2306.00526*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Instruction Set Components

We provide the three key components of the OCR instruction set: the character encoding table (§A.1), the verification step (§A.2), and the JSON output schema (§A.3). The full instruction set is approximately 3,000 words; we include the components most critical for reproducibility.

A.1 Character Encoding Table (excerpt)

The instruction set specifies exact Unicode codepoints and encoding conventions for each archaic character and diacritical mark:

```
ABBREVIATIONS AND EXPANSIONS
- Letters expanded from titla or
  superscript go in PARENTHESES:
  e.g., v"shy(t)ko, ye(d)no
- Titlo mark U+0483 goes INSIDE
  parentheses for abbreviated
  letters: (l"), (ch"), (ts")
- Titlo on sacra nomina stays
  OUTSIDE parentheses on the base
  letter: g"" for gospod'

ARCHAIC CHARACTERS (never modernize)
- omega (U+03C9): ot-ligature as
  omega(t); plain: omega zmiiyu
- yat (U+0463): virimo, zviri
- big yus (U+046B): tu, zovu(t")
- little yus (U+0467): sya, movyachi
- Latin s for zelo: distinct from
  z and dze
- dze (U+0455): distinct from z
  and Latin s

VOWEL LETTERS
- ye (U+0454) and e (U+0435) are
  BOTH valid, often coexist on one
  card. Do NOT default to either.
  Check EVERY instance against card.
- u (U+04AF) and u (U+0443) are
  BOTH valid. Transcribe as written.
```

A.2 Verification Step

After initial transcription, the instruction set requires a line-by-line comparison against the source image, checking for six specific error types:

```
VERIFY BEFORE SAVING:
Re-read the image and compare against
your transcription line by line.
Check for:
- INSERTED LETTERS: characters added
  that are not on the card
- SUBSTITUTED WORDS: a familiar word
  replacing an unfamiliar one
- ye/e CONFUSION: writing ye where
  the card shows e, or vice versa.
  Check EVERY instance.
- ADDED WORD ENDINGS: adding hard
  sign or other characters where
  the card does not have them
- MODERNIZED SPELLING: archaic
  letters silently replaced with
  modern equivalents
```

```
- MISSING CHARACTERS: letters on
  the card that were skipped
```

A.3 JSON Output Schema

Each card is transcribed into a structured JSON record:

```
{
  "filename": "Scan0042.jpg",
  "card_numbers": {
    "primary": "18",
    "secondary": null,
    "tertiary": null,
    "notes": "18 top-left ink"
  },
  "lines": [
    "line 1 of transcription",
    "line 2 of transcription",
    "..."
  ],
  "source": {
    "city": "Lviv",
    "date": "early 17th c.",
    "reference": "Kron. 3 rev."
  },
  "notes": ""
}
```

The `lines` array preserves exact line breaks from the card, enabling alignment between image lines and transcribed text. The `source` object captures provenance metadata written at the bottom of each card. Card numbers track multiple numbering systems (primary/secondary/tertiary) reflecting successive cataloguing efforts.