AI-Powered Knowledge Discovery in the Digital Library of Old Ephemeral Prints: A Case Study

Maciej Ogrodniczuk[0000000234679424] and Dariusz Czerski[0000000230133483]

Institute of Computer Science, Polish Academy of Sciences
maciej.ogrodniczuk@ipipan.waw.pl

Abstract. We investigate how state-of-the-art Large Language Models (LLMs) can unlock new knowledge from the *Cyfrowa Biblioteka Druków Ulotnych* (CBDU)—a digital collection of early-modern Polish ephemeral prints. Our end-to-end pipeline compares three transcription approaches: pure OCR, LLM-based post-correction, and multimodal models. The resulting transcriptions then serve as input for our two main contributions: the automatic extraction of bibliographic metadata and the generation of expert-style historical commentaries. Experiments show a leading multimodal model excels, reducing transcription CER from 33% to 9%, while achieving high F1-scores for publication place (0.85) and date (0.71), and a 2.31/3 mean score for commentaries. We conclude that large multimodal models can serve as effective "digital archivists", enriching historical collections with structured metadata and contextual analysis.

Keywords: digital library · middle Polish · Large Language Models

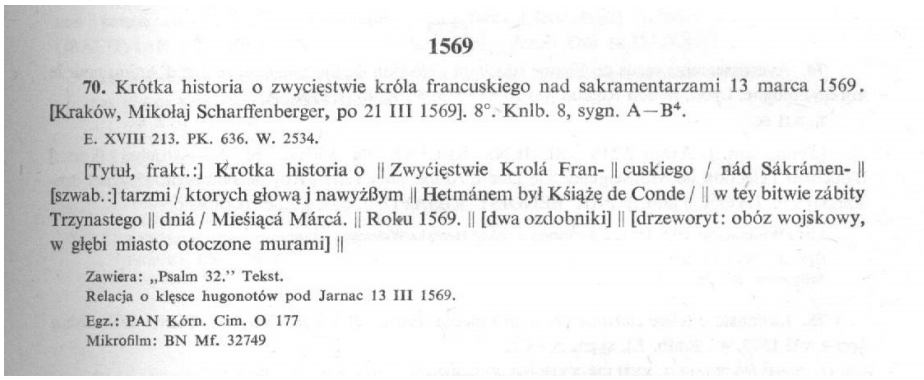
1 Introduction

The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries (Polish: *Cyfrowa Biblioteka Druków Ulotnych polskich i Polski dotyczących z XVI, XVII i XVIII wieku*, abbreviated CBDU[1] [6, 12]) contains approximately 2 000 Polish and Poland-related ephemeral prints (short, disposable, irregular pre-press documents), written in Polish and 10 other languages[2] and dated between 1501 and 1729. The CBDU was manually created in 2009 and has been intensively used by lexicographers [1], corpus linguists [7], media scholars [10], or even developers of text recognition solutions [8].

This library's foundation is a three-volume bibliography by Zawadzki [15–17], a former curator and head of the Microfilm Collection Department at the National Library of Poland. He catalogued and described all surviving items bibliographically and microfilmed many of them. Zawadzki created bibliographic descriptions of the prints containing extensive metadata such as the print title (in modern transcription), issue date and place, printer name, format, volume size, information on bibliographical sources and an exact description of the title

[1] https://cbdu.ijp.pan.pl/
[2] German, Italian, French, Latin, Swedish, Spanish, Dutch, Czech, English and Danish.

page with line breaks, font names, illustrations etc (see Fig. 1). A small number of prints have been also analysed from a historical perspective to provide context for the events they describe.



1569

70. Krótka historia o zwycięstwie króla francuskiego nad sakramentarzami 13 marca 1569. [Kraków, Mikołaj Scharffenberger, po 21 III 1569]. 8°. Knlb. 8, sygn. A—B⁴.

E. XVIII 213. PK. 636. W. 2534.

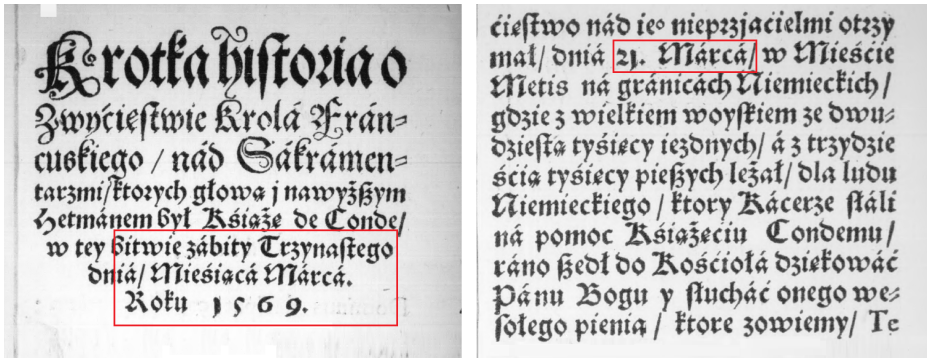
[Tytuł, frakt.:] Krotka historia o || Zwyćięstwie Krolá Fran- || cuskiego / nád Sákrámen- || [szwab.:] tarzmi / ktorych głową j nawyżßym || Hetmánem był Kśiąże de Conde / || w tey bitwie zábity Trzynastego || dniá / Mieśiącá Márcá. || Roku 1569. || [dwa ozdobniki] || [drzeworyt: obóz wojskowy, w głębi miasto otoczone murami] ||

Zawiera: „Psalm 32." Tekst.
Relacja o klęsce hugonotów pod Jarnac 13 III 1569.

Egz.: PAN Kórn. Cim. O 177
Mikrofilm: BN Mf. 32749

Fig. 1. Description of a sample print from Zawadzki's bibliography (see also https://cbdu.ijppan.pl/id/eprint/700/).

When no exact information on metadata such as the publication date was available, it was often reconstructed by looking at the last dated event described in the print. The fact that this type of data has been inferred (rather than being present directly in the document) is indicated by placing the values in square brackets. Fig. 2 illustrates this reconstruction process. The print whose bibliographic description is shown in Fig. 1 contains no publication date, but the title page describes an event (the French king's victory over the Huguenots, French Protestants) dated 13 March 1569 so the print must have been published no earlier than that. However, page 11 of the print contains information about another victory achieved by the same king on 21 March, meaning the publication date was eventually determined to be "after 21 March 1569".



Krotka historia o Zwyćieſtwie Krolá Fráncuskiego / nád Sákrámentarzmi/ktorych głowa j nawyżßym Hetmánem był Kśiąże de Conde/ w tey bitwie zábity Trzynaſtego dniá/ Mieśiącá Márcá. Roku 1569.

cieſtwo nád ie⁹ nieprzyjacielmi otrzymał/ dniá 21. Márcá/ w Mieście Metis ná grániców Niemieckich/ gdzie z wielkiem woyſkiem ze dwudzieſta tyśiecy iezdnych/ á z trzydzieśćia tyśiecy pießych leżał/ dla ludu Niemieckiego/ ktory Kácerze śtáli ná pomoc Kśiążęciu Condemu/ ráno ßedł do Kośćiołá dziękować Pánu Bogu y ſłuchać onego weſołego pienia/ ktore zowiemy/ Te

Fig. 2. Fragments of pages 1 and 11 of print 70 illustrating the process of date reconstruction.

In this paper, we aim to address the following research question: how effectively can modern AI solutions (in this case, multimodal and non-multimodal large language models) be used to automatically process scanned items and discover information relevant to a particular digital library? This can be broken down into three types of task: 1) processing source text; 2) extracting metadata such as publication place or date, either directly from the text or by reasoning through the whole text; and 3) generating descriptions based on the printed text, such as historical comments. The quality of the first two tasks can be evaluated using simple comparisons, whereas the third task requires more sophisticated methods, such as using LLMs as judges. For our case study, we will use CBDU as a source of manually created evaluation data for all three tasks.

2 LLM-powered Processing Pipeline

LLMs can offer unprecedented advancements in processing of historical data, starting from multimodal OCR, through post-correction to Named Entity Recognition [5], just to name a few. In CBDU such models were also previously used for various tasks, including machine translation of Latin interjections with GPT-3 [13] or intralingual diachronic translation of middle-Polish texts into contemporary Polish with GPT-3.5 [9].

Traditional OCR solutions based on deep learning can be applied to historical documents, but achieving high recognition accuracy requires model tuning and large datasets [14]. These approaches are multi-step processes where each step can significantly impact the final result. Modern OCR pipelines typically consist of several key stages:

1. **Preprocessing**, where images are enhanced through noise reduction, contrast adjustment, and deskewing to improve text clarity;
2. **Text detection and segmentation**, where deep learning models identify text regions and separate individual text lines;
3. **Text recognition**, where convolutional neural networks (CNNs) combined with recurrent neural networks (RNNs) or transformer architectures like TrOCR [11] extract character sequences from segmented text regions; and
4. **Post-processing**, where language models and statistical methods correct recognition errors and improve accuracy.

Recent investigations into using LLMs (Large Language Models) for improving OCR recognition have demonstrated substantial ability to recognize text in historical documents [4, 5, 14], though evaluations have primarily used small datasets [5] and not on Polish documents. Multimodal LLMs have shown particular promise in document understanding tasks, offering capabilities that extend beyond traditional OCR to include semantic understanding and contextual analysis.

The use of Large Language Models for OCR recognition and post-processing represents a particularly promising approach for processing historical Polish texts.

2.1 Text Transcription

To transcribe the historical Polish prints we used three strategies: Tesseract 5.5.1 with a Polish language model served as a purely optical-character-recognition baseline; the Tesseract output was post-processed with a text-only LLM pipeline combining Llama-3.3-70B and C4AI Command-A [2]; and a multimodal pipeline in which Llama-4-Scout-17B and Gemini 2.5 Pro [3] processed both the OCR text and the corresponding page image.

For each LLM-based approach we designed a dedicated prompt that asked the model to return two complementary outputs: (i) a *diplomatic transcription* that faithfully preserves the original spelling, punctuation and line breaks, with uncertain characters marked in square brackets; and (ii) a *modern transcription* normalised to contemporary Polish orthography.

While the diplomatic version is essential for detailed philological study, the modern transcription is what truly opens the collection to readers. It enables full-text search with current spelling, feeds later NLP steps such as tokenisers and name-entity recognisers trained on modern Polish, and makes the text instantly readable for historians, linguists and the wider public who are not familiar with early-modern spelling. Creating this reader-friendly version often requires slow manual post-correction, but current LLMs can now produce it *off the shelf*, removing a long-standing bottleneck in digital workflows.

The models were required to respond with a valid JSON object containing only these two fields. The complete prompts used for both text-only and multimodal transcription can be found in Appendix 4.

While we do not formally evaluate the quality of the *modern transcription*, it is a critical intermediate step. By harmonising archaic spellings and normalising grammar to contemporary standards, this output provides a cleaner, more consistent input for the downstream metadata extraction and commentary generation tasks, thereby reducing the risk of error propagation.

2.2 Metadata Discovery

Metadata extraction is a crucial step in digital libraries, as it provides structured information about the documents that can be used for search, analysis and discovery. In this study we explored how modern Large Language Models (LLMs) can be employed to automatically extract such metadata for the CBDU collection. We focused on three key fields: publisher name, place of publication, and date of publication. For the date, we required the models to return values in the canonical format `[after DD roman-month YYYY]`; when an explicit publication date was absent, the earliest date mentioned in the text had to be returned, enclosed in square brackets to mark the inherent uncertainty. We follow the convention introduced in Zawadzki's bibliography, where square brackets enclose any part of the date that has been *inferred* rather than explicitly found in the print. The Polish particle *po* ("after") precedes the reference day and month; typical cases include `[po 3 II] 1586` (where the year is known but the day and month are inferred from the text), `3 II 1586`, `[1610]`, and `[po 2 X 1581]`.

Using the *modern transcription* from the previous stage as input, we evaluated three state-of-the-art LLMs: two text-only engines (C4AI Command-A and Llama-3.3-70B) and one multimodal system (Gemini 2.5 Pro). The full metadata-extraction prompt is provided in Appendix 4.

2.3 Generation of Historical Commentaries

Historical commentaries greatly enhance the value of a digital library by situating each print in its proper temporal and geographical context and by identifying the key actors and events it describes. The creation of comprehensive historical commentaries is a resource-intensive process, demanding both extensive expertise from trained historians and considerable time commitment. We therefore investigated the use of Large Language Models (LLMs) to automate the task, framing commentary generation as a specialised form of abstractive summarisation that benefits from the extensive historical knowledge embedded in contemporary models. Two prompting strategies were devised: a *zero-shot* prompt that requires no in-context examples, and a *five-shot* prompt that provides the model with five reference commentaries for additional guidance. The full text of both prompts can be found in Appendix 4.

As input to the language model we supplied the *modern transcription* generated in the previous stage. Because this version is normalised to contemporary Polish, it provides a substantially cleaner signal than the raw diplomatic transcription. Feeding the LLM with such linguistically harmonised text reduces the cognitive load required to interpret archaic spellings, thereby enabling the model to focus on extracting historical facts and narrative coherence.

3 Evaluation

3.1 Text Transcription

We assessed the quality of automatic transcriptions on a representative subset of 229 Polish ephemeral prints for which human–annotated diplomatic transcriptions were available. Performance was measured with two standard OCR metrics: *Character Error Rate* (CER) and *Word Error Rate* (WER). Both metrics are calculated after normalising the reference and hypothesis strings by (1) removing hyphenated line breaks, (2) stripping superfluous newlines, and (3) collapsing multiple spaces into one. For WER we additionally lowercase all tokens and remove punctuation.

The results in Table 1 show that the purely optical baseline (Tesseract) achieves a CER of $\sim 33\%$ and an extremely high WER of 79%, reflecting the well-known limitations of out-of-the-box OCR on early-modern Polish. Surprisingly, post-processing the Tesseract output with text-only large language models (C4AI Command-A and Llama-3.3-70B) provides almost no improvement, indicating that language-model inflation of errors may offset any contextual correction gains. Introducing image context in multimodal pipelines substantially

Table 1. Transcription accuracy on 229 Polish prints (lower values are better).

System	CER (%)	WER (%)
Pure Tesseract 5.5.1	32.93	79.22
Tesseract + C4AI Command-A (text-only)	33.24	79.12
Tesseract + Llama-3.3-70B (text-only)	33.93	80.18
Tesseract + Llama-4-Scout-17B (multimodal)	29.41	73.47
Tesseract + Gemini 2.5 Pro (multimodal)	**8.89**	**27.23**

improves recognition. Llama-4-Scout-17B reduces both CER by over three percentage points and WER by nearly six percentage points over Tesseract, while Gemini 2.5 Pro cuts the character error rate by a factor of nearly four and more than halves the word error rate. These results underscore the importance of visual grounding for reading degraded historical prints and highlight that very large multimodal language models like Gemini demonstrate strong capabilities for error correction in low-resource, early-modern Polish settings.

3.2 Metadata Discovery

To quantify how accurately large language models can extract bibliographic metadata from early-modern Polish prints, we evaluated three systems—LLAMA-3.3, C4AI Command-A, and Gemini 2.5 Pro on a curated subset of 313 CBDU documents. Only items that already contained at least one of the three target fields (*publisher name*, *place of publication*, *date of publication*) in the ground-truth metadata were included (publisher = 56 documents, place = 205, date = 182). For every field we computed *Precision*, *Recall* and *F1-score* under two matching regimes: *Strict comparison*, which counts a prediction as correct only when it reproduces the ground-truth value exactly and distinguishes between values literally present in the text and those deduced on the basis of the text. *Relaxed comparison*, which ignores that distinction and accepts either form as correct. All metrics were averaged over the set of documents that contain a gold value for the respective field. We do not count false positives for documents where the ground-truth field is empty, as the reference bibliography is known to be incomplete. Penalising a model for extracting a value absent from the ground truth would unfairly punish it for discovering information potentially missed during the original manual annotation.

The results presented in Table 2 reveal three clear trends. First, Gemini 2.5 Pro is the strongest overall performer. Across both matching regimes, it achieves near-perfect recall on publication dates ($\approx 99\%$) and delivers the highest precision, resulting in the best F1-scores for every field. This suggests a clear advantage for larger models when parsing noisy OCR from historical Polish documents.

Second, the matching regime primarily affects date extraction. Relaxed comparison boosts F1-scores by 6–9 points across all systems, indicating that many errors stem from differences in data extraction confidence levels rather than ac-

Table 2. Field-level metadata extraction accuracy (higher is better).

Model	Comp.	Field	Precision	Recall	F1	Docs
Llama-3.3-70B	Strict	Date of publication	0.169	0.086	0.114	182
		Place of publication	0.791	0.727	0.758	205
		Publisher	0.583	0.512	0.546	56
	Relaxed	Date of publication	0.292	0.140	0.189	182
		Place of publication	0.823	0.734	0.776	205
		Publisher	0.583	0.512	0.546	56
C4AI Command-A	Strict	Date of publication	0.198	0.282	0.233	182
		Place of publication	**0.851**	0.516	0.642	205
		Publisher	0.704	0.396	0.507	56
	Relaxed	Date of publication	0.298	0.371	0.330	182
		Place of publication	**0.868**	0.521	0.651	205
		Publisher	0.704	0.396	0.507	56
Gemini 2.5 Pro	Strict	Date of publication	**0.448**	**0.988**	**0.616**	182
		Place of publication	0.837	**0.847**	**0.842**	205
		Publisher	**0.771**	**0.563**	**0.651**	56
	Relaxed	Date of publication	**0.553**	**0.990**	**0.709**	182
		Place of publication	0.854	**0.849**	**0.852**	205
		Publisher	**0.771**	**0.563**	**0.651**	56

tual incorrect values. Publisher and place metrics are mostly unaffected due to less ambiguity in the ground truth.

Third, the text-only models show complementary strengths. C4AI Command-A achieves higher precision on place and publisher extraction, while Llama-3.3-70B has stronger recall, leading to better F1-scores for those fields. For dates, C4AI-Command-A is superior. These diverse error profiles suggest that an ensemble strategy could be effective, especially for the low-resource publisher and date fields where no single model excels at both precision and recall.

Overall, this study confirms that modern LLMs can recover bibliographic metadata from early-modern Polish prints with promising accuracy. However, future work will require targeted methods to normalise date formats and disambiguate publisher names.

3.3 Generation of Historical Commentaries

To quantify the quality of the automatically generated historical commentaries we evaluated 93 prints for which expert gold-standard commentaries are available. Following the *LLM-as-a-judge* paradigm [18], we asked `gpt-4o-2024-05-13` (knowledge cut-off: June 2024) to assign an ordinal score to each system output by comparing it with the ground truth according to the rubric: **1** – no overlapping historical information (or wrong language); **2** – partial overlap (more than half of the relevant facts); **3** – high similarity ($\geq$ 90% of the facts). The full prompt can be found in Appendix 4. The evaluation was carried out in a pointwise manner, i.e. each candidate commentary was judged independently without exposing the model to alternative hypotheses.

Table 3. Detailed breakdown of historical-commentary quality. Counts denote how many of the 93 test prints were assigned a given judge score. Higher mean values are better.

Engine	Prompting	Mean	#1	#2	#3
Gemini 2.5 Pro	zero-shot	2.258	8	53	32
Gemini 2.5 Pro	few-shot	**2.312**	9	46	38
C4AI Command-A	zero-shot	1.978	15	65	13
C4AI Command-A	few-shot	2.097	12	60	21
Llama-3.3-70B	zero-shot	1.892	24	55	14
Llama-3.3-70B	few-shot	1.946	18	62	13

The results in Table 3 reveal four clear trends. Gemini 2.5 Pro consistently outperforms the other engines, indicating that larger models translate their superior language understanding into higher factual coverage when generating historical summaries. Providing *few-shot* examples yields a modest but systematic improvement for every engine (0.05–0.12 absolute points), confirming that in-context examples help steer the models towards the desired factual structure. Smaller models (C4AI Command-A, Llama-3.3-70B) lag behind Gemini by about 0.2–0.4 points, suggesting that they struggle to capture subtle historical details present in early-modern Polish prints. Overall, the evaluation supports the use of advanced LLMs with few-shot prompting when high-precision historical commentary is required. Notably, Gemini also exhibits the lowest occurrence of low-quality outputs: only 8/93 commentaries (8.6 %) are scored 1 in the zero-shot condition and 9/93 (9.7 %) in the few-shot variant, whereas competing systems range between 13–26 %. At the same time, few-shot training boosts Gemini's share of perfect commentaries (score 3) from 34 % to 41 %, underscoring its capacity to translate minimal guidance into substantial quality gains.

4 Conclusions and Future Work

Our study demonstrated that multimodal LLMs can effectively process and enhance historical prints by improving transcription, normalizing archaic text, and generating expert-quality historical commentaries. Compared to traditional OCR and text-only LLMs, multimodal approaches significantly reduce errors and produce more comprehensive factual coverage, emphasizing the importance of visual context. These findings suggest that advanced LLMs can act as "digital archivists," streamlining labor-intensive archival tasks while enriching digital collections with accessible transcriptions and contextual insights.

This work paves the way for two future scenarios. First, LLMs can be used to supplement existing digital libraries like CBDU by generating structured metadata and contextual information—such as relationships between items—that were beyond the scope of the original project. A more general application involves using a multimodal LLM as a self-contained system to create a digital library from scratch, automating the entire archival workflow from scanned image to a living, searchable knowledge base.

References

1. Bronikowska, R., Majdak, M., Wieczorek, A., Żółtak, M.: The Electronic Dictionary of the 17th- and 18th-century Polish – Towards the Open Formula Asset of the Historical Vocabulary. In: Gavriilidou, Z., Mitsiaki, M., Fliatouras, A. (eds.) Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion. vol. I, pp. 471–475. Democritus University of Thrace (2020), https://euralex.org/publications/the-electronic-dictionary-of-the-17th-and-18th-century-polish-towards-the-formula-asset-of-the-historical-vocabulary/
2. Cohere, T., Aakanksha, Ahmadian, A., Ahmed, M., Alammar, J., Alnumay, Y., Althammer, S., Arkhangorodsky, A., Aryabumi, V., Aumiller, D., Avalos, R., Aviv, Z., Bae, S., Baji, S., Barbet, A., Bartolo, M., Bebensee, B., Beladia, N., Beller-Morales, W., Bérard, A., Berneshawi, A., Bialas, A., Blunsom, P., Bobkin, M., Bongale, A., Braun, S., Brunet, M., Cahyawijaya, S., Cairuz, D., Campos, J.A., Cao, C., Cao, K., Castagné, R., Cendrero, J., Currie, L.C., Chandak, Y., Chang, D., Chatziveroglou, G., Chen, H., Cheng, C., Chevalier, A., Chiu, J.T., Cho, E., Choi, E., Choi, E., Chung, T., Cirik, V., Cismaru, A., Clavier, P., Conklin, H., Crawhall-Stein, L., Crouse, D., Cruz-Salinas, A.F., Cyrus, B., D'souza, D., Dalla-Torre, H., Dang, J., Darling, W., Domingues, O.D., Dash, S., Debugne, A., Dehaze, T., Desai, S., Devassy, J., Dholakia, R., Duffy, K., Edalati, A., Eldeib, A., Elkady, A., Elsharkawy, S., Ergün, I., Ermis, B., Fadaee, M., Fan, B., Fayoux, L., Flet-Berliac, Y., Frosst, N., Gallé, M., Galuba, W., Garg, U., Geist, M., Azar, M.G., Goldfarb-Tarrant, S., Goldsack, T., Gomez, A., Gonzaga, V.M., Govindarajan, N., Govindassamy, M., Grinsztajn, N., Gritsch, N., Gu, P., Guo, S., Haefeli, K., Hajjar, R., Hawes, T., He, J., Hofstätter, S., Hong, S., Hooker, S., Hosking, T., Howe, S., Hu, E., Huang, R., Jain, H., Jain, R., Jakobi, N., Jenkins, M., Jordan, J., Joshi, D., Jung, J., Kalyanpur, T., Kamalakara, S.R., Kedrzycki, J., Keskin, G., Kim, E., Kim, J., Ko, W.Y., Kocmi, T., Kozakov, M., Kryściński, W., Jain, A.K., Teru, K.K., Land, S., Lasby, M., Lasche, O., Lee, J., Lewis, P., Li, J., Li, J., Lin, H., Locatelli, A., Luong, K., Ma, R., Mach, L., Machado, M., Magbitang, J., Lopez, B.M., Mann, A., Marchisio, K., Markham, O., Matton, A., McKinney, A., McLoughlin, D., Mokry, J., Morisot, A., Moulder, A., Moynehan, H., Mozes, M., Muppalla, V., Murakhovska, L., Nagarajan, H., Nandula, A., Nasir, H., Nehra, S., Netto-Rosen, J., Ohashi, D., Owers-Bardsley, J., Ozuzu, J., Padilla, D., Park, G., Passaglia, S., Pekmez, J., Penstone, L., Piktus, A., Ploeg, C., Poulton, A., Qi, Y., Raghvendra, S., Ramos, M., Ranjan, E., Richemond, P., Robert-Michon, C., Rodriguez, A., Roy, S., Ruis, L., Rust, L., Sachan, A., Salamanca, A., Saravanakumar, K.K., Satyakam, I., Sebag, A.S., Sen, P., Sepehri, S., Seshadri, P., Shen, Y., Sherborne, T., Shi, S.C., Shivaprasad, S., Shmyhlo, V., Shrinivason, A., Shteinbuk, I., Shukayev, A., Simard, M., Snyder, E., Spataru, A., Spooner, V., Starostina, T., Strub, F., Su, Y., Sun, J., Talupuru, D., Tarassov, E., Tommasone, E., Tracey, J., Trend, B., Tumer, E., Üstün, A., Venkitesh, B., Venuto, D., Verga, P., Voisin, M., Wang, A., Wang, D., Wang, S., Wen, E., White, N., Willman, J., Winkels, M., Xia, C., Xie, J., Xu, M., Yang, B., Yi-Chern, T., Zhang, I., Zhao, Z., Zhao, Z.: Command A: An Enterprise-Ready Large Language Model (2025), arXiv:2504.00698
3. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities (2025), arXiv:2507.06261

4. Do, T., Tran, D.P., Vo, A., Kim, D.: Reference-Based Post-OCR Processing with LLM for Precise Diacritic Text in Historical Document Recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 27951–27959 (2025). https://doi.org/10.1609/aaai.v39i27.35012
5. Greif, G., Griesshaber, N., Greif, R.: Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents (2025), arXiv:2504.00414
6. Gruszczyński, W., Ogrodniczuk, M.: Digital Library 2.0: Source of Knowledge and Research Collaboration Platform. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1649–1653. European Language Resources Association (ELRA), Reykjavik, Iceland (2014), https://aclanthology.org/L14-1164/
7. Gruszczyński, W., Adamiec, D., Bronikowska, R., Kieraś, W., Modrzejewski, E., Wieczorek, A., Woliński, M.: The Electronic Corpus of 17th- and 18th-century Polish Texts. Language Resources and Evaluation **56**(1), 309–332 (2022). https://doi.org/10.1007/s10579-021-09549-1
8. Idziak, J., Šeundefineda, A., Woźniak, M., Leśniak, A., Byszuk, J., Eder, M.: Scalable Handwritten Text Recognition System for Lexicographic Sources of Under-Resourced Languages and Alphabets. In: Computational Science – ICCS 2021: 21st International Conference. Proceedings, Part I. pp. 137–150. Springer-Verlag, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-77961-0_13
9. Klamra, C., Kryńska, K., Ogrodniczuk, M.: Evaluating the Use of Generative LLMs for Intralingual Diachronic Translation of Middle-Polish Texts into Contemporary Polish. In: Goh, D.H., Chen, S.J., Tuarob, S. (eds.) Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration. ICADL 2023. pp. 18–27. No. 14457 in Lecture Notes in Computer Science, Springer Nature Singapore, Singapore (2023). https://doi.org/10.1007/978-981-99-8085-7_2
10. Kolasa, W.M.: Kierunki badań nad prasą polską najstarszej doby (1501–1729) (*Directions in research of the oldest Polish press (1501–1729)*), in Polish. Studia Medioznawcze (3 (50)), 65–80 (2012)
11. Liu, M., Li, T., He, G., Li, H., Wang, Y., Wang, J., Qiao, Y., Liu, J., Li, Z., Tang, J.: TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models (2021). https://doi.org/10.48550/arXiv.2109.10282
12. Ogrodniczuk, M., Gruszczyński, W.: Digital Library of Poland-related Old Ephemeral Prints: Preserving Multilingual Cultural Heritage. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage. pp. 27–33. Hissar, Bulgaria (2011), http://www.aclweb.org/anthology/W11-4105
13. Ogrodniczuk, M., Kryńska, K.: Evaluating Machine Translation of Latin Interjections in the Digital Library of Polish and Poland-related News Pamphlets. In: Tseng, Y.H., Katsurai, M., Nguyen, H.N. (eds.) From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries. ICADL 2022. pp. 430–439. No. 13636 in Lecture Notes in Computer Science, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-21756-2_34
14. Patel, D.: Comparing Traditional OCR with Generative AI-Assisted OCR: Advancements and Applications. International Journal of Science and Research (IJSR) **14**, 347–351 (2025). https://doi.org/10.21275/SR25603211507
15. Zawadzki, K.: Gazety ulotne polskie i Polski dotyczące z XVI, XVII i XVIII wieku. Bibliografia. Tom I: 1514–1661 (En. *Polish and Poland-related Ephemeral Prints*

from the 16th-18th Centuries. Bibliography. Volume 1: 1514–1661), in Polish. National Ossoliński Institute, Polish Academy of Sciences, Wrocław (1977)
16. Zawadzki, K.: Gazety ulotne polskie i Polski dotyczące z XVI, XVII i XVIII wieku. Bibliografia. Tom II: 1662–1728 (En. *Polish and Poland-related Ephemeral Prints from the 16th-18th Centuries*. Bibliography. Volume 2: 1662–1728), in Polish. National Ossoliński Institute, Polish Academy of Sciences, Wrocław (1984)
17. Zawadzki, K.: Gazety ulotne polskie i Polski dotyczące z XVI, XVII i XVIII wieku. Bibliografia. Tom III: 1501–1725 (En. *Polish and Poland-related Ephemeral Prints from the 16th-18th Centuries*. Bibliography. Volume 3: 1501–1725), in Polish. National Ossoliński Institute, Polish Academy of Sciences, Wrocław (1990)
18. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 46595–46623 (2023). https://doi.org/10.5555/3666122.3668142

Appendix A: LLM Prompts

A.1 Text-only (non-multimodal) pipeline

```
You are an expert historian and philologist specialised in early-modern
printed sources. I will provide you with 1) the raw OCR-recognised
text of a single page (which can contain errors and uses historical
orthography).
Sources are in Polish language.
Please analyse input and return *only* a valid JSON object with
two properties:
    - "diplomatic_transcription" - a faithful, line-break preserving
      transcription that imitates the original spelling and punctuation
      as closely as possible; use square brackets [] to mark uncertain
      characters.
    - "modern_transcription" - the same text normalised to contemporary
      orthography and punctuation (expand contractions, modernise
      letters, correct obvious typos, but keep the semantics).
Do not wrap the JSON in Markdown fences; output plain JSON only.
```

A.2 Multimodal pipeline

```
You are an expert historian and philologist specialised in early-modern
printed sources. I will provide you with 1) the raw OCR-recognised
text of a single page (which can contain errors and uses historical
orthography) and 2) the scan of that very page as an image.
Sources are in Polish language.
Please analyse both inputs and return *only* a valid JSON object with
two properties:
    - "diplomatic_transcription" - a faithful, line-break preserving
      transcription that imitates the original spelling and punctuation
      as closely as possible; use square brackets [] to mark uncertain
      characters.
```

```
    - "modern_transcription" - the same text normalised to contemporary
      orthography and punctuation (expand contractions, modernise
      letters, correct obvious typos, but keep the semantics).
Do not wrap the JSON in Markdown fences; output plain JSON only.
```

A.3 Historical commentary prompts

```
0-shot Prompt:

You are an expert historical analyst. Your task is to extract a concise
↪  historical commentary from the following Polish-language old print
↪  text. Focus on identifying and clearly describing the following key
↪  elements:

Period of Time  When did the event or situation described take place?
↪  (Provide specific years or general historical period if not
↪  explicitly stated.)

Place  Where did the event occur? (Indicate the city, region, or
↪  country.)

Event  What is the main historical event or development described in the
↪  print?

Main Participants  Who were the key individuals, social groups, or
↪  institutions involved?

Write the commentary in an informative, academic tone, suitable for
↪  historical research or archival documentation.

Important: The original text is written in Polish. Interpret accordingly.
Important: Provide additional historical context for the events described
↪  in the text.

The output text should be short, no more than 10 sentences.

Input (Polish old print text):

VERY IMPORTANT: keep the output text as one CONTINUOUS text in POLISH
↪  LANGUAGE.

{INPUT_TEXT}


5-shot Prompt:

You are an expert historical analyst. Your task is to extract a concise
↪  historical commentary from the following Polish-language old print
↪  text. Focus on identifying and clearly describing the following key
↪  elements:
```

```
Period of Time  When did the event or situation described take place?
↪  (Provide specific years or general historical period if not
↪  explicitly stated.)

Place  Where did the event occur? (Indicate the city, region, or
↪  country.)

Event  What is the main historical event or development described in the
↪  print?

Main Participants  Who were the key individuals, social groups, or
↪  institutions involved?

Write the commentary in an informative, academic tone, suitable for
↪  historical research or archival documentation.

Important: The original text is written in Polish. Interpret accordingly.
Important: Provide additional historical context for the events described
↪  in the text.

The output text should be short, no more than 10 sentences.

Take the below examples of historical commentaries as a reference:

<example 1>
<example 2>
<example 3>
<example 4>
<example 5>

VERY IMPORTANT: keep the output text as one CONTINUOUS text in POLISH
↪  LANGUAGE.

Input (Polish old print text):

{INPUT_TEXT}
```

A.4 Metadata extraction prompt

```
You are an expert bibliographer and historian specializing in
↪  early-modern printed sources. Your task is to extract key publication
↪  metadata from the following Polish-language old print text.

Please analyze the text and extract the following information:

1. **Publisher Name** - The name of the person, institution, or company
↪  responsible for publishing the work. If the publisher's name is not
↪  mentioned in the text, leave the field blank.
```

```
2. **Place of Publication** - The city, town, or location where the work
↪  was published. If the place of publication is not mentioned in the
↪  text, leave the field blank.
3. **Date of Publication** - The date when the work was published, in the
↪  format [after DD roman-month YYYY]. If no exact publication date is
↪  given in the text, use the earliest date mentioned as the publication
↪  date. Enclose any uncertain part of the date in square brackets [ ].

Important guidelines:
- Look for explicit statements about publication details in the text
- Consider historical context and common publication practices of the
↪  period
- If information is not explicitly stated, indicate "Not specified" or
↪  "Unknown"
- For dates, provide the most specific information available (year,
↪  approximate period, etc.)
- For places, provide the most specific location mentioned
- For publishers, include full names when available

Output your findings in the following JSON format:

{
    "publisher": "Name of the publisher or 'Not specified'",
    "place_of_publication": "City/town where published or 'Not
    ↪  specified'",
    "date_of_publication": "Year or date when published or 'Not
    ↪  specified'",
    "year": "Year of events described in the print or 'Not specified'"
}

Do not wrap the JSON in markdown fences; output plain JSON only.
OUTPUT OTHER THEN JSON WILL CAUSE TOTAL SYSTEM FAILURE!

Input (Polish old print text):

{INPUT_TEXT}
```

A.5 Historical commentary judging prompt

```
Compare the two historical commentaries provided below. The first is the
↪  ground truth, and the second is the automatically extracted version.
↪  Your task is to evaluate how much historical information from the
↪  ground truth is present in the automatic extraction.

Use the following 3-point scale for your judgment:

1. No common information  The automatic extraction contains none of the
↪  historical information found in the ground truth. Returns true if the
↪  extracted commentary is in a language other than Polish.
```

```
2. Partial information  The automatic extraction includes more than half
↪  of the historical information from the ground truth.

3. High similarity  The automatic extraction contains more than 90% of
↪  the historical information found in the ground truth.

The commentaries are written in Polish.

Extracted Commentary:

{EXTRACTED}

Ground Truth Commentary:

{GROUND_TRUTH}

Justification:

Final score: <score number 1-3>
```