Digitizing the Chicago Assyrian Dictionary: Accelerating Document Analysis with VLMs and HPC Infrastructure

Himanshi Yadav*, Christian Borgen†, Hervé Reculeau†, Foy Scalf†, and Ross Hyman*
*Research Computing Center
†Institute for the Study of Ancient Cultures
University of Chicago, Chicago, Illinois, 60637

***Abstract*—The Chicago Assyrian Dictionary (CAD), completed in 21 plus volumes over 90 years (1921–2011), is a foundational Akkadian lexical reference locked in print and print-derived images. Manual text generation from dictionary scanned images requires domain expertise and is time-consuming. We present a generalizable pipeline that converts dense mixed-script dictionary pages into accurate UTF-8-encoded structured data. The HPC-enabled pipeline combines image preprocessing and layout segmentation with Vision Language Model (VLM)–based Optical Character Recognition (OCR), prompt engineering, and post-processing. In few-shot experiments on the gold standard dataset, Claude Opus 4.6 achieved a 1.45% CER and 5.75% WER on the training set (for optimal prompt design and selection) and 1.49% CER and 4.73% WER on the validation set. Using the pipeline, running on CPU compute nodes, we extracted raw text in less than a week, with domain expert verification underway to finalize the scholarly dataset for human-environment history research.**

I. Motivation

Hundreds of thousands of Akkadian cuneiform tablets, originating over thousands of years (3000-100 BCE) have been found throughout Mesopotamia. The Chicago Assyrian Dictionary (CAD), completed in 21+ volumes over 90 years (1921–2011), is a foundational lexical reference of the Akkadian language [1]. The CAD is a monumental achievement of scholarship from which enormous insight can be drawn about ancient Mesopotamian societies and cultures, but the dictionary is publicly available only in print form and as scanned images of its printed pages, thereby limiting searching, indexing, extracting text, or further computational analysis. General-purpose OCR tools enable conversion from images to text but are not designed for the typographic and linguistic density of Assyriological lexicography. Vision-language models (VLMs) have proven proficient on multi-language OCR tasks [2], [3]. An objective of this work is to consider the role of university HPCs in machine learning research, with increased research access to frontier machine learning capabilities, and stretch the HPC compute towards domain expert aligned corpus development tasks. More specifically, we explore the capabilities of open-weight and closed-weight VLMs by benchmarking on the OCR task to transcribe print CAD page images. Further, we provide a pipeline design to leverage HPC infrastructure to enable multi-level domain expert data alignment, to convert the CAD image corpus to a relational text database.

II. Approach

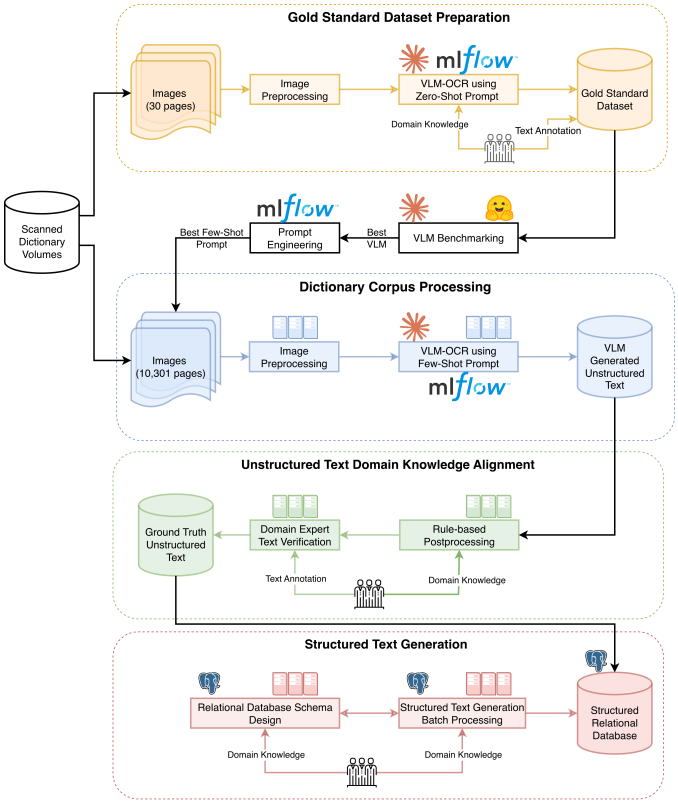


Fig. 1. HPC-enabled pipeline for converting scanned print dictionary volumes to a structured relational database.

We present a pipeline (Fig. 1) for converting scanned dictionary images into a structured relational database. The pipeline consists of the following:

1) **Gold Standard Dataset Preparation** VLM-assisted image-to-text generation and domain expert text alignment to prepare a gold standard dataset. The gold standard development set comprises 20 dictionary pages, and the validation set has 10 dictionary pages.
2) **VLM Benchmarking** Best VLM selection by benchmarking open-weight and closed-weight VLMs on the gold standard dataset. We locally deployed Hugging

Face open-weight models [4], namely, Qwen3-VL-2B [5] and OlmOCR-2-7B [6], and employed API access for closed-weight proprietary Claude models.

3) **Prompt Engineering** Optimal prompt design and selection, using the best-performing VLM, on the gold standard dataset with MLflow-facilitated [7] experiment tracking.
4) **Rule-based Postprocessing** Rule-based postprocessing of VLM-generated dictionary text corpus, using domain expert-identified VLM errors and edge cases.
5) **Domain Expert Text Alignment** Domain experts and annotators make corrections to VLM-generated dictionary text corpus at scale.
6) **Structured Relational Database Generation** Relational database schema design and unstructured to structured text conversion with domain expert inputs.

The key challenges encountered during this work are a lack of ground truth data, high lexicographic density of CAD text, a lack of VLM interpretability, and optimal prompting under data scarcity conditions.

III. Results

Claude Opus 4.6 produced the best zero-shot performance on 30 pages of the gold standard dataset with 2.40% CER and 6.80% WER (Tab. I). In few-shot experiments, it achieved a 1.45% CER and 5.75% WER on the development set (split into sets for prompt engineering) and 1.49% CER and 4.73% WER on the validation set.

TABLE I
Zero-Shot VLM Benchmarks: Average Character Error Rate and Word Error Rate (in %) on Gold Standard Dataset

Model	CER	WER
Qwen3-VL-2B	45.02	68.02
OlmOCR-2-7B	8.25	21.62
Sonnet 4.5	7.25	21.95
Opus 4.6	**2.40**	**6.80**

Each CAD dictionary volume is submitted as a separate SLURM batch job with job-level logging, and Claude API requests are made from CPU cluster compute nodes. MLflow is leveraged for experiment tracking at the volume level and run tracking at the page level. To address the high text density per page, the dictionary page images are split into halves locally before constructing the API request. A transcription prompt configuration, followed by a merge prompt configuration, is employed to convert the dictionary page images to markdown page text. Markdown preserves the inherent structure-aware text font styling of the print CAD and interfaces easily with the frontier model API output formats. Figure 2 shows the input token and output token costs per CAD dictionary volume. The input token costs are double that of the output token costs due to the use of two prompt configurations.

By leveraging CPU cluster compute nodes and VLMs for the OCR task, we transcribed the dictionary page images in less than a week (Fig. 3). Domain expert alignment of VLM-generated text is underway to arrive at the ground truth unstructured text corpus.

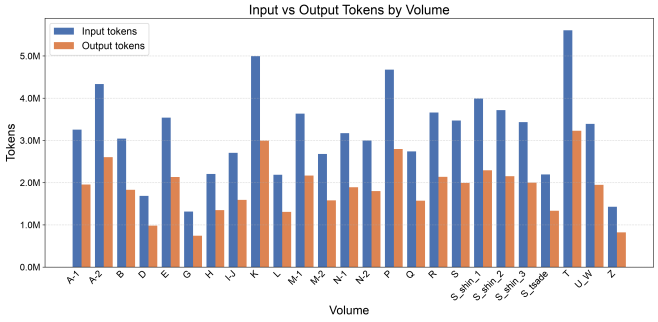


Fig. 2. Input token and output token costs (in counts) by CAD dictionary volume.

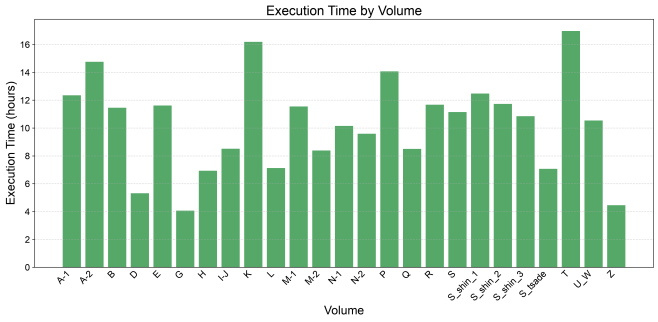


Fig. 3. VLM-assisted OCR execution time (in hours) by CAD dictionary volume.

IV. Conclusion and Future Work

This work demonstrates that VLMs, enabled by HPC infrastructure, can accelerate scholarly dataset development for human-environment history research. Domain expert verification of postprocessed VLM-generated text is underway. On comparing Volume Q's postprocessed text against domain expert aligned ground truth text, we achieved a 4.04% CER and 3.60% WER. We continue efforts on domain expert aligned relational database development.

Acknowledgment

This work is supported by the AI Pillar Initiative at the University of Chicago and the Yelda Family Fund at the Institute for the Study of Ancient Cultures.

References

[1] M. Roth, "How we wrote the chicago assyrian dictionary," *JNES*, vol. 69, pp. 1–21, 2010.
[2] Y. Li, G. Yang, H. Liu, B. Wang, and C. Zhang, "dots.ocr: Multilingual document layout parsing in a single vision-language model," 2025. [Online]. Available: https://arxiv.org/abs/2512.02498
[3] Z. Zhang, H. Liu, S. Liang, Y. Zhang, Y. Xiang, J. Liu, T. Sun, M. Lin, Y. Zhang, C. Zhou, T. Gao, C. Cui, Y. Liu, D. Yu, and Y. Ma, "Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training," 2026. [Online]. Available: https://arxiv.org/abs/2606.03264

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/

[5] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2511.21631

[6] J. Poznanski, L. Soldaini, and K. Lo, "olmocr 2: Unit test rewards for document ocr," 2025. [Online]. Available: https://arxiv.org/abs/2510.19817

[7] M. A. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the Machine Learning Lifecycle with MLflow," *IEEE Data Eng. Bull.*, vol. 41, pp. 39–45, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:83459546