



CuneiML: A Cuneiform Dataset for Machine Learning

DANLU CHEN 

ADITI AGARWAL 

TAYLOR BERG-KIRKPATRICK 

JACOBO MYERSTON 

*Author affiliations can be found in the back matter of this article

COLLECTION:
REPRESENTING THE
ANCIENT WORLD
THROUGH DATA

DATA PAPER

ubiquity press

ABSTRACT

The cuneiform writing system holds a vast reservoir of ancient literature, encompassing over 3000 years of history. Originating around the mid-fourth millennium BCE and enduring until the late first millennium BCE, cuneiform writing spans various genres such as administrative, legal, medical, and scientific documents, among others. This article introduces a curated dataset, CuneiML, featuring 38,947 high-resolution 2D photos of Sumerian and Akkadian cuneiform tablets, accompanied by their cuneiform Unicode transcriptions, transliterations, lineart, and metadata. This dataset aims to support the development of machine learning tools for processing and analyzing Sumerian and Akkadian cuneiform artifacts – e.g. for automatically classifying genre, provenance, or period from unannotated tablet images. Thus, CuneiML is designed with consistency of format as a primary concern. Specifically, CuneiML is a result of meticulously preprocessing, segmenting, filtering, and re-transliterating data that is available online in the Cuneiform Digital Library Initiative (CDLI) collection.

CORRESPONDING AUTHOR: Danlu Chen

Computer Science and
Engineering, UC San Diego,
La Jolla, US

dac013@ucsd.edu

KEYWORDS:

cuneiform; machine learning;
computational paleography;
image processing

TO CITE THIS ARTICLE:

Chen, D., Agarwal, A., Berg-Kirkpatrick, T., & Myerston, J. (2023). CuneiML: A Cuneiform Dataset for Machine Learning. *Journal of Open Humanities Data*, 9: 30, pp. 1–9. DOI: <https://doi.org/10.5334/johd.151>

1 INTRODUCTION

In this article we present a curated dataset of 38,947 2D photographs of Sumerian and Akkadian cuneiform tablets with their accompanying transcriptions in cuneiform Unicode – as well as lineart, transliterations, and metadata specifying attributes like period and genre. In contrast to the data provided by digital libraries which offer general access to cuneiform texts, our dataset was envisioned from the very beginning for machine learning with an emphasis on consistency of format. Therefore, we developed our dataset with strict preprocessing and filtering criteria and present preliminary baseline experiments for three classification tasks supported by our data: period, provenance, and genre prediction, conditioned on major face cutouts from tablet photographs.

The CuneiML dataset was produced by processing photographs and transliterations available online in the Cuneiform Digital Library Initiative (CDLI) (Englund et al., 2023). This library gives access to 56,694 photographs of inscribed objects classified by time period, genre, provenance, and museum collection. Current digitized cuneiform archives like CDLI were designed as portals where experts can consult photographs of inscribed objects (tablets, seals, inscriptions, etc.), transliterations, dictionaries, and other working tools. Although the CDLI is an extraordinary resource which has proven to be of invaluable use for Assyriologists, it offers its data in a format not suitable for machine learning experiments. CDLI photographs are of varied quality: some are high-resolution while others do not meet the minimum requirements for machine learning tasks. In addition, CDLI images are composite; this means they contain multiple perspectives of the same object: front, back and sides of tablets (see Figure 1). Another issue is that the transliterations of Sumerian and Akkadian that accompany CDLI images are missing their rendering into cuneiform Unicode. These aspects make the CDLI data unsuitable for machine learning. Thus, with machine learning in mind, we have meticulously filtered and processed the CDLI corpus, isolating the most salient fragments of 38,947 high resolution composite images and have tokenized and converted the Latin transliterations of Sumerian and Akkadian into cuneiform Unicode. This latter part of the dataset retains the original polysemy of the cuneiform sign, a feature that neural network architectures like transformers are capable of capturing (Gari Soler & Apidianaki, 2021).

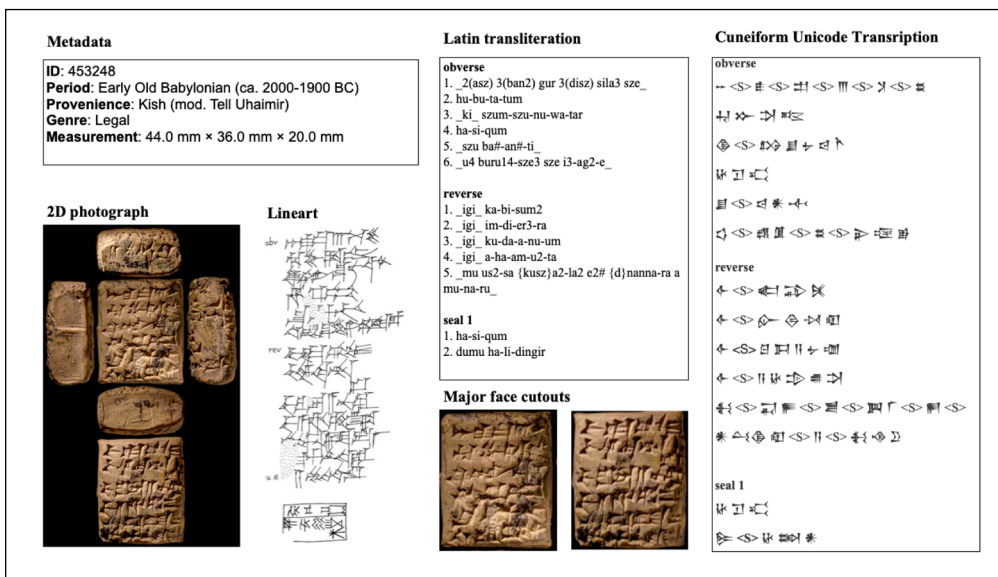


Figure 1 An overview of CuneiML. An example tablet of ID 453248 with multi-modal data: (1) **Metadata** consist of time period, provenience, genre and measurement. (2) High-resolution **2d photograph** of 6 faces. (3) **Lineart** from paleographers. (4) **Latin transliteration** directly downloaded from CDLI. (5) **Cuneiform Unicode transcription** we automatically converted from the Latin transliteration. (6) **Major face cutouts** automatically processed from the 2d photograph.

Although 3D scans are preferable to 2D photographs, to our knowledge there is only one existing Open Access 3D dataset of cuneiform tablets, which is limited in size and the historical periods. The Hilprecht – Heidelberg Cuneiform Benchmark Dataset for the Hilprecht Collection (HeiCuBeDa), contains 3D scans of only 1,977 tablets, which are limited to merely four historical periods, namely, Ed IIIb (ca. 2500-2340 BCE), UR III (2100-2000 BCE), Old Assyrian (ca. 1950-1850 BCE), and Old Babylonian (1900-1600 BCE) (Bogacz & Mara, 2020). In an ideal world, all cuneiform tablets in museums would be 3D scanned, but such a scenario is not foreseeable in the near future. Based on existing photographs that have been collected by museums and scholars, our data provide a 20 times larger number of tablets covering almost the entire history

of cuneiform writing, as well as richer metadata information, linearts (hand drawings made by modern scholars), and transcriptions into cuneiform Unicode.

Our dataset supports the development of a variety of machine learning tools – for example, the training and evaluation of automatic classifiers for predicting period, genre, or provenance from an artifact’s photograph, its Unicode transcription, or a transliteration of that transcription; but also, development of – for example – end-to-end automatic transcription systems from lineart or photograph (Gutherz, Gordin, Sáenz, Levy, & Berant, 2023). Next, we provide summary info for our dataset, followed by a description of how it was collected and processed. Finally, we include initial experiments with baseline classifiers on three classification tasks supported by our dataset.

2 DATASET DESCRIPTION

As we stated in the introduction, our aim is to curate a dataset that can support the development of novel machine learning tools for cuneiform. Our data consist of composite 2d photographs of tablets, as well as their major face cutouts, lineart, transliteration, transcription into cuneiform Unicode, and metadata. An example is shown in Figure 1. We also plot the histograms for time period (Figure 2), genre and provenience.

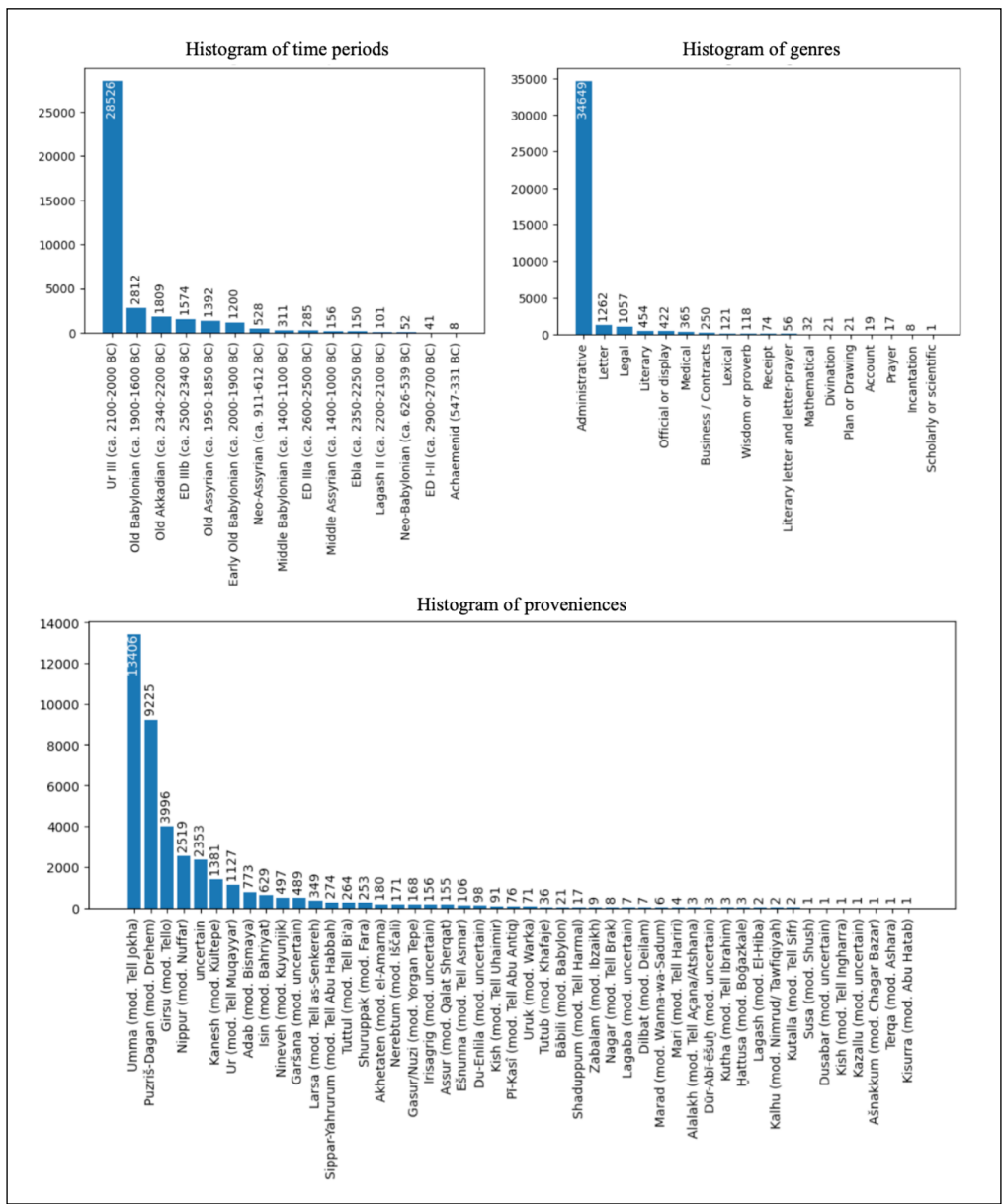


Figure 2 Number of tablets by metadata attributes: time period, genre, and provenience.

2.1 SUMMARY

Below, we provide a brief summary of metadata for CuneiML:

Object name CuneiML_v1.0.tar.gz, tran.

Format names and versions JPEG and JSON.

Creation dates 2023-09-01

Dataset creators Danlu Chen, Aditi Agarwal, Taylor Berg-Kirkpatrick, Jacobo Myerston.

Language Sumerian and Akkadian.

License CC BY-NC 4.0.

Repository name <https://doi.org/10.5281/zenodo.8307503>

Publication date 2023-09-01

3 METHODS

Our processing methods for dataset creation break down into several phases. First, we systematically scrape composite 2D photographs, transliterations, and artifact metadata for all artifacts represented in CDLI. Next, we process the composite 2D photographs in order to split them into images of individual tablet faces. Finally, we develop a set of de-transliteration rules for converting each transliteration into Unicode. Throughout this process, we automatically filter out unusual and rare forms of artifacts that could introduce spurious correlations into the prediction tasks supported by our dataset. In the next sections, we describe each phase of the pipeline in more detail. For implementation details, please checkout the code repository <https://github.com/taineleau/CuneiML>.

3.1 DOWNLOADING THE METADATA FROM CDLI

We used the CDLI Github repository¹ to get the public catalog data containing a list of P-numbers for all tablets. “P-number” is a 6 digit unique identifier prefixed with the letter P, used to uniquely identify a tablet by the CDLI initiative. Using the P-number for every tablet we then crawl 2D images,² lineart images³ and metadata⁴ including transliteration from CDLI. We gathered a total of 133,923 tablets from CDLI, of which 56,694 come with a 2D scan image and 52,637 come with a lineart image.

3.2 DATA SAMPLING AND FILTERING

Most 2D photographs from CDLI are high-resolution color images showing six faces of a single tablet. Since our goal is to create a dataset for training and evaluating machine learning systems, consistency is paramount: if non-standard examples are included (e.g. black and white images when the majority of the dataset is in full color), machine learning systems may learn to leverage these features as false predictors (e.g. if most black and white images tend to come from the same period due to how they were collected, a classifier will learn to depend on this spurious correlation). Thus, we filtered out non-standard and low-quality and images according to several conditions:

- The image is black-and-white.
- The resolution of image is lower than 100*100 px.
- The tablet is in poor condition, e.g. too many fragments or barely readable cuneiform.
- The image does not contain a well-defined major face.

1 <https://github.com/cdli-gh/data>.

2 <https://cdli.mpiwg-berlin.mpg.de/dl/photo/P000001.jpg>.

3 https://cdli.mpiwg-berlin.mpg.de/dl/lineart/P000001_1.jpg.

4 <https://cdli.mpiwg-berlin.mpg.de/artifacts/1/json>.

We also filtered out artifact type such as cone, cylinder and prism and only keep entries whose artifact type is tablet. After this processing, we eventually have 38,947 tablets with high quality 2D images.

3.3 CUTTING OUT THE MAJOR FACES

An additional issue for machine learning development arises from the variability of the imaging setup used to capture the raw composite photographs from CDLI. While the majority of composite photographs consist of six tablet faces, arranged in a fixed layout from a consistent camera angle, these properties vary to some extent based on when and where imaging was performed. In order to increase consistency, and therefore reduce the danger of overfitting to false correlations that are dependent on the imaging process itself, we systematically extract cutouts of the major tablet face in each composite and include this as an additional, more consistent, photographic representation for machine learning systems. Specifically, we build and test three different computer vision methods to segment and obtain individual faces of each tablet. As a way of validating and producing final extractions, we reconcile differences between the bounding boxes each system produces by computing the area of their overlap. When the area is large, the methods are in agreement and our output is reliably high-quality. The three methods are described briefly as follows.

- 1. Connected component segmentation.** We first convert the images into black and white where the background is black. This is a classical rule-based segmentation that clusters the adjacent pixels with the same color. We use OpenCV's implementation `cv.connectedComponents()`.
- 2. Watershed segmentation.** We first convert the images to grayscale. The watershed algorithm views a grayscale image as a topographic surface where high intensities denote hills while low intensities denote valleys. Each valley is labelled with a different color of water. As the water rises, unknown pixels will be colored and therefore clustered. We use OpenCV's implementation `cv.watershed()`.
- 3. SegmentAnything.** This is the state-of-the-art general segmentation algorithm using neural networks. We use the official toolkit⁵ (Kirillov et al., 2023) with default model weights to obtain cutouts.

Quality checking We automatically cut the images using three methods and only keep tablet cutouts whose overlapping area is larger than 90%. We sampled 100 images randomly to validate the cutouts; 97% met our quality requirements. Figure 3 shows a sample of 20 major cutouts produced by our algorithm.



Figure 3 A random sample of 20 major face cutouts.

3.4 CONVERTING TRANSLITERATION TO CUNEIFORM UNICODE

CDLI and ORACC offer transliterations of cuneiform tablets, which are sometimes but not always accompanied by their photographs and linearts. The transliteration standard used in both projects is called ATF and is explained in detail in the ORACC's website.⁶ Transliteration is the process of transcribing cuneiform signs into the Latin alphabet, using conventions which have varied over time and can take particular shapes according to various Assyriological projects. But, despite its possible inconsistencies, transliteration has played

⁵ <https://github.com/facebookresearch/segment-anything>.

⁶ <http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html>.

a crucial role in making Akkadian and Sumerian more accessible to the non-specialist; it has also facilitated the creation of dictionaries and critical editions. From the point of view of data processing, transliteration is also important because it reveals how modern scholars read certain signs that allow multiple interpretations. In this sense, transliteration is a form of disambiguation. Take for example the sign \star that can read as “sky” or “god” but can also be interpreted as the syllables *an* or *il*. This issue of interpretation occurs with many cuneiform signs that need to be disambiguated so that modern editors can stabilize what seems to them the most plausible reading of a text. A final example may serve to further illustrate this point. In the well-known *Epic of Creation* or *Enūma eliš*, the mother of the gods’ name is often spelled with the signs TI and GÉME, a combination of signs that is usually transliterated as *Ti-amat* and rendered into English as Tiamat. Now, this transliteration somewhat conceals that the goddess is the Sea, something that can be expressed more directly if one transliterates TI GÉME as *ti-amtu*, the “sea” in Akkadian. Thus, transliteration implies a reduction of possible choices which were present for an ancient audience, but which are concealed to modern readers that use latinized editions of cuneiform texts.

One possible issue if we use the transliteration directly for machine learning, is **circular reasoning**. Given that the transliteration itself might exhibit bias towards specific time periods and other attributes – e.g., an expert’s approach to transliterating a tablet is already influenced by preconceived notions about its time period. Thus, as an additional layer in our dataset, we produce and provide cuneiform Unicode conversions of the original Latin transliterations. Specifically, we follow the ATF convention to remove some of the editorial marks, tokenize, and map the transliteration into machine-readable cuneiform Unicode format. We use **cuneifyplus**⁷ to map the Latin transliteration to cuneiform Unicode. If a latinized sign is not processed, we then query eBL’s sign list⁸ to obtain the cuneiform Unicode. We briefly describe the rules here.

1. **Uncertainty.** The query (?) placed after a grapheme indicates uncertainty and the asterisk (*) indicates a collated reading. We remove the marks but keep the grapheme by default.
2. **Breakage.** The \$ sign represents breakage, sometimes also indicating how many lines are broken. If the number is recorded, we insert the same number of <LB>. E.g. 2 lines broken → <BREAK><LB><BREAK><LB>. Moreover, the annotation [...] indicates missing signs. We also insert a special token <BREAK> to indicate the missing content.
3. **Compound words.** We remove the markers of compound words. E.g. |SU.KUR| → su-kur.
4. **Reading.** sudx(|SU.KUR|) means the reading is sudx, while the signs are su-kur; we remove the reading and only keep the actual signs for tokenization.

Quality checking

We downloaded and extracted a dataset with human annotated transliteration-transcription pairs from the Akkademia project⁹ to use as a validation reference for our method. We take 2,719 lines of the human-labeled Latin transliteration/cuneiform Unicode transcription pairs. We run our program to tokenize and convert the transliteration into the Unicode transcription and obtain 99% character accuracy against the reference.

4 POTENTIAL USAGE AND TASKS

As described above, each cuneiform tablet in CuneiML comes with multiple layers of information across several modalities. The potential usages of this dataset for machine learning development can be roughly split into unimodal and multi-modal tasks (the possible inputs and outputs from

7 <https://github.com/tpgillam/cuneifyplus>.

8 <https://www.ebl.lmu.de/signs>.

9 <https://github.com/gaigurtherz/Akkademia>.

cuneiML are summarized in [Table 1](#)). Beyond the more standard classification tasks like period, genre, and provenance prediction, we suggest several additional examples of potential tasks that our dataset can support. This list is not intended to be exhaustive.

TASK NAME	INPUT	OUTPUT
Language Modeling	(4)(5)	(4)(5)
Transliteration	(5)	(4)
Lineart generation	(2)(6)	(3)
Attribute prediction	(2)(3)(4)(5)(6)	(1)
Sign identification	(2)(3)(6)	(5)

Table 1 Task summary with possible input and output pairs. (1) Metadata consist of time period, provenience, genre and measurement. (2) High-resolution 2d photograph of 6 faces. (3) Lineart from paleographers. (4) Latin transliteration (5) Cuneiform Unicode transcription. (6) Major face cutouts.

UNIMODAL TASKS

- **Language modeling.** One of the most popular and broadly useful machine learning applications is to train a language model on text in a given domain. Language models trained on our dataset could be used to encode cuneiform Unicode for further processing and analysis, or as a generative prior in related downstream tasks like transcription and restoration ([Assael et al., 2022](#); [Lazar et al., 2021](#)).
- **Transliteration.** As mentioned above, there are multiple ways to transliterate the same cuneiform sign sequence. [Gordin et al. \(2020\)](#) proposed several models, including HMMs and LSTMs, to automatically transliterate and segment Unicode cuneiform glyphs. Our dataset could be used as further training or validation data for this task.
- **Lineart generation.** Analogously, in the image modality, there are potential use cases for automatically “translating” photographic representations into lineart, which potentially increases the readability of tablets to scholars. This task is structurally similar to image generation tasks in the broader field of computer vision. Following similar techniques, CuneiML could be used to train neural models ([Isola, Zhu, Zhou, & Efros, 2017](#); [Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022](#)) capable of accurate lineart generation conditioned on a tablet image.

Multi-modal tasks

- **Attribute prediction.** Our dataset supports training and evaluating classifiers for predicting metadata based on images or lineart. The attributes in the metadata include geographical, genre, and chronological attribution ([Bogacz & Mara, 2020](#)).
- **Sign identification / automatic transcription.** A useful, but particularly challenging task that our data supports is automatic transcription of tablet images or lineart into cuneiform Unicode text. There is very little text line annotation data for cuneiform tablets. Even with line-level annotations, the task is much harder than documents written on paper. Recently, new page-level end-to-end OCR systems ([Coquenot Chatelain, & Paquet, 2023](#)) have been developed that are capable of high-accuracy transcription of more modern languages without line-level annotation. The lineart-cuneiform Unicode parallel data presented in our dataset is an ideal testbed for extending these techniques to more ancient languages.

In the following section, we present preliminary results on attribute prediction tasks using major face images, cuneiform Unicode and Latin transliteration in order to demonstrate a specific use case of our dataset for machine learning.

5 PRELIMINARY EXPERIMENTS WITH ATTRIBUTE PREDICTION

We analyze the task of attribute prediction for cuneiform and present the results on an image classification baseline using deep neural networks. We take three different types of attributes (time period, provenance, and genre) and treat each separately as a target output for an automatic classifier. As shown in [Figure 2](#), the distribution of these attributes are imbalanced

and long-tailed; therefore, we discard classes whose number of examples are less than 50. We then split the data randomly into training, validation and testing sets with a ratio of {0.9, 0.05, 0.05}. In all cases, our image classifier is a pretrained version of ResNet-101 that we continue to optimize on our training set. For the textual features, we train a two-layer LSTMs from scratch.

RESULT AND ANALYSIS

Table 2 shows the result of attribute prediction using three different type of input as features. We can see that the image baseline model achieves reasonable accuracy on all three types of attributes. The preliminary results demonstrate the effectiveness of our data for training machine learning systems that make predictions based in tablet images and further underscore the difficulty of provenience and genre attribution. Note that the major face cutout features seem to have the overall best performance, but it is possible that lighting and camera configurations influence the classification of a tablet. Furthermore, label imbalance and the distribution shift between the training and testing sets remain significant challenges in cuneiformML.

	IMAGE	UNICODE	TRANS.	# OF CLASSES
Time period	97.66	90.50	87.17	14
Provenience	85.72	61.71	68.60	25
Genre	89.00	81.50	86.21	12

Table 2 Summary of test accuracy for attribute prediction using different features.


Further research and analysis are necessary to assess the reliability of the predicted results.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Danlu Chen  orcid.org/0000-0002-8582-8151
 Computer Science and Engineering, UC San Diego, La Jolla, US

Aditi Agarwal  orcid.org/0009-0005-2878-1018
 Computer Science and Engineering, UC San Diego, La Jolla, US

Taylor Berg-Kirkpatrick  orcid.org/0000-0002-1283-4075
 Computer Science and Engineering, UC San Diego, La Jolla, US

Jacobo Myerston  orcid.org/0000-0003-0865-9494
 Literature Department, UC San Diego, La Jolla, US

REFERENCES

- Assael, Y., Sommerschild, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., ... de Freitas, N.** (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, 603, 280–283. DOI: <https://doi.org/10.1038/s41586-022-04448-z>
- Bogacz, B., & Mara, H.** (2020). Period Classification of 3D Cuneiform Tablets with Geometric Neural Networks. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 246–251). DOI: <https://doi.org/10.1109/ICFHR2020.2020.00053>
- Coquenat, D., Chatelain, C., & Paquet, T.** (2023). Dan: a segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7), 8227–8243. DOI: <https://doi.org/10.1109/TPAMI.2023.3235826>
- Englund, R. K., Damerow, P., Pagé-Perron, E., Dahl, J. L., Lafont, B., & Renn, J.** (2023). *Cuneiform Digital Library Initiative*. Retrieved 2023-08-22, from <https://cdli.mpiwg-berlin.mpg.de/>
- Garí Soler, A., & Apidianaki, M.** (2021). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9, 825–844. DOI: https://doi.org/10.1162/tacl_a_00400
- Gordin, S., Guthertz, G., Elazary, A., Romach, A., Jiménez, E., Berant, J., & Cohen, Y.** (2020). Reading Akkadian cuneiform using natural language processing. *PLOS ONE*, 15(10), e0240511. DOI: <https://doi.org/10.1371/journal.pone.0240511>
- Guthertz, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J.** (2023). Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5), pgad096. DOI: <https://doi.org/10.1093/pnasnexus/pgad096>

- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A.** (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134). DOI: <https://doi.org/10.1109/CVPR.2017.632>
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R.** (2023). Segment anything. *arXiv preprint arXiv:2304.02643*. DOI: <https://doi.org/10.48550/arXiv.2304.02643>
- Lazar, K., Saret, B., Yehudai, A., Horowitz, W., Wasserman, N., & Stanovsky, G.** (2021). Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4682–4691). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2021.emnlp-main.384>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B.** (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695). DOI: <https://doi.org/10.1109/CVPR52688.2022.01042>

TO CITE THIS ARTICLE:

Chen, D., Agarwal, A., Berg-Kirkpatrick, T., & Myerston, J. (2023). CuneiML: A Cuneiform Dataset for Machine Learning. *Journal of Open Humanities Data*, 9: 30, pp. 1–9. DOI: <https://doi.org/10.5334/johd.151>

Submitted: 02 September 2023

Accepted: 17 October 2023

Published: 06 December 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.