

# Img2SumGlyphs: Transformer-based OCR of Sumerian Cuneiform

Cole Simmons  
Stanford University  
coles@stanford.edu

## Abstract

*The corpus of Sumerian texts represents the birth of writing, a written tradition spanning 3,000 years, and a rare union of textual and material artifact. More than 120,000 have been uncovered to date; these texts provide a ground-level view of the most foundational chapters in the story of human civilization. However, learning to read Sumerian takes years of specialized training. With so many texts and so few experts capable of reading them, many have yet to be read. Moreover, only a handful have accessible translations, preventing non-specialists from engaging with the vast majority of the corpus. A set of deep learning-based tools could enable Sumerologists to parse and translate the corpus quickly, accurately, and at scale.*

*In this paper, I introduce SumTablets\_Photos<sup>1</sup>, the first dataset of image–Unicode glyph sequence pairs. Then, I use this dataset to train a model on an optical character recognition (OCR) task: given an input image, it autoregressively generates the represented Unicode glyphs. This model, Img2SumGlyphs, combines a fine-tuned version of the vision transformer (ViT) encoder used in TrOCR and XLM-R decoder in an encoder–decoder architecture. Img2SumGlyphs is the first Sumerian OCR model and establishes a baseline performance of 35.41 character error rate (CER) on a held-out test set to be improved upon in future work. Altogether, this work sets the foundation for a new approach to Sumerology—one which promises to soon make the earliest writing accessible to everyone.*

## 1. Introduction

Sumerian is the earliest attested written language. Originating in southern Mesopotamia (modern-day Iraq south of Baghdad) around 2900 BCE, Sumerian continued to be written for another 3,000 years. In the third millennium BCE, it was primarily used for admin-



Figure 1. An administrative Sumerian cuneiform tablet from Shuruppak (mod. Tell Fara), dated to the Early Dynastic IIIa period (ca. 2500 BCE). [4]

istration and royal inscriptions. And although many believe that Sumerian went extinct as a spoken language around 2000 BCE, Babylonian scribes nonetheless continued to use Sumerian as the preferred written language of literature, liturgy, mathematics, science, and other cultic or scholastic contexts. Today we call their writing system “cuneiform” (from the Latin *cuneus* “wedge” due to the wedge-like components of glyphs). Cuneiform was originally devised to encode Sumerian but was later adapted to encode more than a dozen languages throughout the Near East. Because cuneiform texts were written on durable materials like clay and stone, they have survived to the present in tremendous quantity [9], with more than 120,000 of these being Sumerian.

Yet despite the abundance of Sumerian texts, accessing their contents remains a significant challenge. The struggle to fully decipher the language continues to this day, and even for those who have spent years learning Sumerian, reading is slow and laborious. Sumerologists read and collaborate in the medium of *transliteration*, a conventional system for rendering glyph readings phonetically in the Latin alphabet [18]. As a result many transliterations are available online, but ex-

<sup>1</sup>Published on Hugging Face

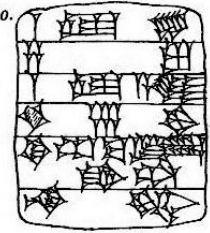
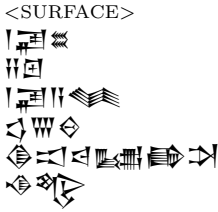
ID	Image (Input)	Glyphs (Target)	Period	Genre
P105897		<p>&lt;SURFACE&gt;</p> 	Ur III	Administrative

Table 1. A sample row from *SumTablets\_Photos*. The model takes as input an image, encodes it using a ViT, autoregressively generates glyphs using an XLM-R model, and the generations are compared against the true targets. Glyph sequences also contain structural information, retained through the use of extra-semantic special tokens such as <SURFACE> and \n. While the dataset also includes photos, I limit the scope of model training and evaluation here to only lineart images.

ceedingly few translations are available. The lack of available translations prevents non-experts, from the general public to generalist historians, from engaging directly with these texts.

Deep learning tools have the potential to address these challenges by enabling Sumerologists to transliterate and translate new tablets far more rapidly than ever before, focusing on verifying the generations rather than creating each new transliteration and translation from scratch. Furthermore, these language models would enable a positive flywheel of data quality: Experts can efficiently review discrepancies between the source and generation, as well as low-confidence predictions; retraining the model with improved data and regenerating predictions leads to a new set of items to review, and so on—rapidly improving the quality and consistency of the data. However, progress on applying NLP to the sub-tasks in reading Sumerian has been limited by the lack of standardized, structured data.

The most plentiful data for tablets are transliterations (published by Sumerologists) and images (published by museums). This data is not immediately suited to any OCR or NLP task: Because each glyph has dozens of possible readings depending on the context, the most sensible and observable process is to have separate models and datasets dedicated to glyph extraction, transliteration, and translation. But although cuneiform began to be added to Unicode in 2006, it has not been widely utilized because of inefficient input. No textual representation of the glyphs previously existed.

To address this issue, I create and publicly release *SumTablets\_Photos*, the first dataset that pairs images of tablets with Unicode representations of the glyphs. It comprises **79,633 image–glyph pairs**, and a total of **6,592,453** glyphs. Of these pairs, 36,891 are lineart images (as in Table 1) and 42,742 are photos (as in

Figure 1). I create this dataset by cleaning and standardizing available transliterations and working backwards to each reading’s source glyph. This dataset enables the development of optical character recognition (OCR) models that learn to extract Unicode representations of glyphs present in images, a necessary first step for subsequent transliteration and translation models.

I use *SumTablets\_Photos* to train an OCR model, *Img2SumGlyphs*, which combines TrOCR’s vision transformer encoder with a pretrained XLM-R decoder. *Img2SumGlyphs* achieves a state-of-the-art character error rate (CER) score of 35.41. With the release of this dataset, definition of the task, and establishment of baseline results, this work sets the foundation for applying recent advances in deep learning to the world’s most ancient corpus.

## 2. Related Work

### 2.1. Machine learning and Sumerian

To the best of my knowledge, this work is the first to attempt any sort of Sumerian OCR system. Prior applications of machine learning techniques to Sumerian have trained models to extract Part of Speech (POS) and Named Entity Recognition (NER) information from Sumerian cuneiform [3], translate Sumerian to English [19], and align Sumerian transliterations with images of tablets [6]. Deep learning models have found success in restoring fragmented ancient texts [2], with some models being applied to predict missing Sumerian transliterations from Neo-Babylonian texts [8]. Only one other work [10] has utilized cuneiform Unicode, doing so to develop a transliteration model for Akkadian. Jauhainen et. al [13] build a dataset for language identification from images of cuneiform glyphs, but do not attempt further glyph identification.



## 2.2. Vision transformers

My Sumerian OCR model utilizes a vision transformer (ViT) for the encoder, which has recently emerged as a powerful architecture for learning effective image representations. Transformers [20], originally developed for NLP, eschew the recurrence and convolutions that were previously ubiquitous in neural networks for sequential data. Instead, they rely on attention mechanisms that model interactions between elements of the input sequence. Dosovitskiy et al. [7] showed that transformers can also be applied to image recognition by splitting the image into a sequence of patches and providing the sequence of linear embeddings of these patches as input to a transformer encoder. ViTs have since achieved state-of-the-art results on many computer vision benchmarks [11].

Most importantly for this work, Li et al. [15] developed TrOCR for handwritten and printed text recognition, combining a ViT encoder initialized from a pretrained BEiT, the weights of the decoder initialized from a pretrained RoBERTa, the cross-attention weights randomly initialized, and then fine-tuning on millions of real or synthesized examples of handwriting. Their results outperformed prior methods for handwriting recognition.

## 2.3. Multilingual modelling

Sumerian is a difficult language to computationally model, as it is both a language isolate and is low-resource. However, large pre-trained multilingual models can learn powerful internal representations of general language features, transferring these learnings to quickly adapt to languages like Sumerian. The efficacy of cross-lingual representations in models such as XLM-R [5], mBART [16], m-T5 [16], and BLOOM [14] is demonstrated by their performance in zero- and few-shot cross-lingual benchmarks such as XTREME [12] and MEGA [1].

## 3. Methods

### 3.1. Model Architecture

To perform optical character recognition (OCR) on images of Sumerian cuneiform tablets, I train a sequence-to-sequence model called *Img2SumGlyphs* that combines a vision transformer (ViT) encoder with an XLM-R decoder. The model architecture based on TrOCR [15], a transformer-based encoder-decoder OCR model that has achieved state-of-the-art results on handwritten and printed text recognition benchmarks. The encoder is initialized from a BEiT vision transformer pre-trained on ImageNet. Vision transformers [7] have recently emerged as a powerful alterna-

tive to the traditional approach of using convolutional neural networks for image recognition tasks. A ViT takes as input an image of size  $3 \times H \times W$  and splits it into a sequence of fixed-size patches. These patches are linearly embedded, supplemented with positional embeddings, and then fed into a standard transformer encoder architecture. The encoder outputs a sequence of image patch embeddings. Concretely, the ViT encoder used in *Img2SumGlyphs* splits the image into patches of  $16 \times 16$  pixels, and the encoder has 12 layers with hidden dimension 768.

The decoder is an XLM-R transformer model [?] pre-trained on a large multilingual corpus spanning 100 languages. XLM-R extends the multilingual BERT model to incorporate more languages and a larger training dataset, achieving state-of-the-art cross-lingual performance. Although Sumerian is a language isolate, it shares independent grammatical features with many languages that are included in XLM-R’s pretraining dataset. For example, Sumerian is agglutinative like Turkish, and has a split-ergative alignment like Basque. Both of these languages—and other relevant ones—are included in the pre-training dataset. Utilizing these representations is crucial for learning the patterns between glyphs. At each decoding step  $t$ , the XLM-R decoder takes the ViT encoder outputs and the embedding of the previous output glyph  $g_{t-1}$  and predicts a probability distribution over the current glyph  $g_t$ :

$$g_t = \text{softmax}(\text{XLM-R}(\text{ViT}(x), g_{t-1}))$$

where  $x$  is the input image. The decoder has 12 layers, 12 attention heads, and an intermediate dimensionality of 3072. The model is trained end-to-end with a standard cross-entropy loss to maximize the log-likelihood of the ground truth glyph sequence  $G = (g_1, \dots, g_T)$  conditioned on the input image:

$$\mathcal{L} = - \sum_{t=1}^T \log p(g_t | g_{<t}, x)$$

I use Hugging Face libraries for dataset management, model instantiation, and training. Before training end-to-end, I initialize a pretrained XLM-R model and fine-tune it on the glyph data with a causal language modelling (CLM) objective. This both facilitates a warm-start where the representations are more tightly aligned with the patterns in Sumerian, and it allows me to take advantage of a fair amount of glyph data for which there are no associated images and thus will not be seen during end-to-end training. I also utilize a custom glyph tokenizer with a vocabulary size of 632 tokens (621 unique Unicode glyphs plus eleven special tokens). Resizing the embeddings of the XLM-R model drops

the number of trainable parameters from 306M to 86M. I initially train the decoder with all layers frozen except for the embedding layer so that the embeddings may adjust appropriately without disrupting the internal representations.

I then initialize a pretrained TrOCR model “microsoft/trocr-base-handwritten” and replace the default decoder with the one pretrained on glyphs. This combination allows the model to build upon both the visual representations learned by the ViT during pre-training on a large dataset of images and the multilingual and language-specific knowledge captured by XLM-R, adapting both components to the specific domain of recognizing Sumerian cuneiform.

The input images are resized to a fixed size of  $384 \times 384$  pixels and normalized before being fed into the ViT. Before resizing the images, I pad either left and right sides or the top and bottom sides, depending on whether the image is portrait (most common) or landscape, respectively. To mitigate overfitting and create more generalizable learnings, I augment the dataset by duplicating each image in the training data set, and randomly rotating the duplicated image within a range of -10 to 10 degrees, and randomly alter the hue, saturation, and lightness by a factor of 20%.

The ground truth glyph sequences contain not only the cuneiform glyphs but also special tokens containing structural information (e.g., <SURFACE> to indicate the start of a new surface and \n to indicate a line break breaks). The model autoregressively generates the glyphs one at a time, conditioned on the input image and all previously generated glyphs until generating the end-of-sequence (EOS) token.

### 3.2. Training

The *Img2SumGlyphs* model is trained on a single NVIDIA A100 SXM GPU with 80 GB memory. I use the AdamW optimizer [17], which extends the commonly used Adam algorithm with a decoupled weight decay regularization, which has been shown to improve training stability and generalization.

For regularization, I apply dropout with probability 0.1 in the XLM-R decoder. Dropout randomly zeros out a fraction of the activations during training, which helps prevent overfitting and improve generalization to unseen data. The learning rate is linearly warmed up over the first 10% of training steps to stabilize the early phases of optimization, and then linearly decayed to zero over the remaining steps. This learning rate schedule allows the model to quickly converge to a good initialization and then fine-tune it with increasingly smaller updates. Finally, as the encoder-decoder began to overfit at the end, I add weight decay factors

of  $1e-5$  and  $1e-4$  (in the final two runs respectively), which mitigated it somewhat but did not prevent overfitting.

I split the *SumTablets\_Photos* dataset into train, validation, and test sets with a ratio of 90/5/5. Since this split was performed before limiting to the linear images only and before dropping some examples, this ratio may not hold exactly in the final dataset. The model is trained on the train set, with the validation loss calculated every 100 steps. After each run, the model with the lowest validation loss is saved.

### 3.3. Evaluation

The trained *Img2SumGlyphs* model is evaluated on the test set using character error rate (CER), a standard metric for assessing the accuracy of OCR systems. CER measures the edit distance between the predicted and ground truth glyph sequences, normalized by the number of glyphs in the ground truth sequence:

$$\text{CER} = \frac{\text{EditDistance}(\text{predicted}, \text{ground truth})}{\text{len}(\text{ground truth})}$$

The edit distance is calculated as the minimum number of single-character insertions, deletions, and substitutions required to transform the predicted sequence into the ground truth sequence. CER ranges from 0 to 1, with 0 indicating a perfect match and 1 indicating that the predicted and ground truth sequences have no overlap. I report the average CER across all examples in the test set, as well as a breakdown by period and genre.

This work establishes a strong foundation and baseline for future research on Sumerian cuneiform OCR. The model architecture combining a pre-trained vision transformer encoder and multilingual text decoder is well-suited to this challenging task, operating on complex images of an ancient, non-Latin script. Successfully training and evaluating the model on the new *SumTablets\_Photos* dataset demonstrates the feasibility and promise of this approach.

However, there remain many opportunities for further improvement. The dataset could be expanded with additional sources of cuneiform images and transliterations, especially those that take advantage of the three-dimensional nature of the texts. More advanced data augmentation techniques could be applied during training to improve robustness. Segmentation would help the model learn more direct relationships.

Nonetheless, this work takes a key first step in tackling Sumerian OCR with deep learning. By open-sourcing the dataset and baseline model, I hope to encourage further research and collaboration on this challenge.

## 4. Dataset and Features

### 4.1. Creating Unicode representations of tablets

In publishing a transliteration, a Sumerologist states how they think a tablet should be read; many such transliterations have been published online. But because Sumerologists are reading directly from either the physical text or an image, no digital representation of the original text’s glyphs is recorded. Although most cuneiform glyphs have now been added to Unicode<sup>2</sup>, no dataset of Sumerian cuneiform tablets represented in Unicode currently exists, barring the development of OCR systems.

To construct *SumTablets\_Photos*, I begin with the data provided by ePSD2/Oracc via JSON files<sup>3</sup>: meta-data and transliterations for 91,606 texts. These transliterations were produced by dozens of research groups over decades of changing conventions and evolving knowledge of Sumerian vocabulary and grammar; they also contain extensive (but not useful for our purposes) embedded ASCII annotation. Most importantly, the data in Oracc do not contain one key piece of information: a representation of the original glyphs from which the transliterations were generated. I first perform extensive cleaning and standardizing the Oracc transliterations. Then, because although each glyph has multiple possible readings, each reading is backed by a single glyph, I am able to work backwards to a parallel representation of the source glyphs. The resulting dataset of glyph–transliteration pairs comprises 91,606 tablets and approximately 7,000,000 glyphs.<sup>4</sup>

### 4.2. Pairing Unicode with tablet images

The glyphs are not inscribed, but impressed; reading them often relies on holding the object in your hand and rotating it, taking advantage of light and shadow to give the characters contrast. For this reason, and because the advent of Assyriology predates high-quality photography, many inscriptions have been published in the “lineart” form (as in Figure 5).

With this on hand, this data was easily joined with the photos and lineart available on <https://cdli.mpiwg-berlin.mpg.de/>.

Cuneiform tablets are three-dimensional objects, so images of tablets are taken from all angles in order to fully capture what is present. However, training a sequence-to-sequence model on these images would be

<sup>2</sup>All online Sumerian data aggregation and collaboration was limited to ASCII for more than a decade: The first cuneiform was added to Unicode in 2006.

<sup>3</sup><https://oracc.museum.upenn.edu/epsd2/json>

<sup>4</sup>I have also published this dataset to Hugging Face: <https://huggingface.co/datasets/colesimmons/SumTablets>

Period	Train	Val	Test
Ur III	24,664	1,417	1,420
Old Akkadian	1,780	102	99
Early Dynastic IIIb	1,671	91	103
Early Dynastic IIIa	370	20	16
Old Babylonian	288	20	17
Lagash II	172	11	7
Early Dynastic I-II	55	1	4
Unknown	4	-	-
Neo-Babylonian	2	-	-
Middle Babylonian	2	-	-
<b>Total</b>	<b>29,008</b>	<b>1,662</b>	<b>1,666</b>

Genre	Train	Val	Test
Administrative	27,443	1,557	1,579
Royal Inscription	514	30	34
Letter	402	30	16
Literary	336	24	21
Legal	283	15	16
Liturgy	10	1	-
Lexical	9	-	-
Math/Science	5	4	-
Unknown	6	1	-
<b>Total</b>	<b>29,008</b>	<b>1,662</b>	<b>1,666</b>

Table 2. Composition by period and tablets of the used subset of *SumTablets\_Photos*. These counts are for lineart images only and are after filtering out tablets with extreme aspect ratios or glyph sequences above the maximum length of 256.

difficult, as the model would have to learn that lines can be started on one surface and wrap onto another. For that reason, I will begin by working only with the lineart representations.

### 4.3. Filtering, processing, and augmentation

I filter out any image that has an aspect ratio below 0.3 or above 1.2, as when the image is padded into a square resized, the glyphs become too small. Furthermore, I filtered out examples that had glyph sequences above my set maximum of 256.

The input images are resized to a fixed size of  $384 \times 384$  pixels and normalized before being fed into the ViT. Before doing so, I pad each image into a square. Padding is applied to either left and right sides or the top and bottom sides, depending on whether the image is portrait (most common) or landscape, respectively. To mitigate overfitting and create more generalizable learnings, I augment the dataset by duplicating each image in the training data set, and randomly ro-

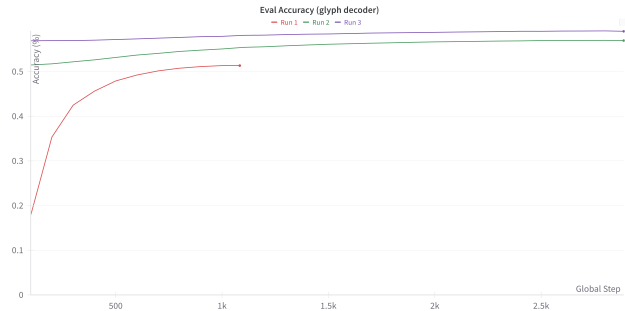


Figure 2. Evaluation accuracy of the glyph decoder during fine-tuning.

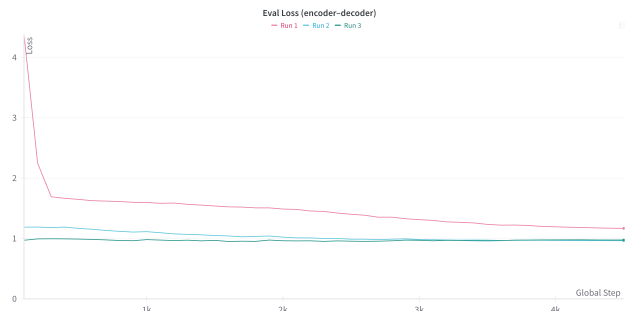


Figure 3. Evaluation loss of the encoder-decoder during fine-tuning.

tating the duplicated image within a range of -10 to 10 degrees, and randomly alter the hue, saturation, and lightness by a factor of 20%.

## 5. Experiments

I first initialize a pretrained XLM-R model and fine-tune it on the glyph data with a causal language modelling (CLM) objective. I initially train the decoder with all layers frozen except for the embedding layer so that the embeddings may adjust appropriately without disrupting the internal representations. For the second run, I unfreeze the final six layers and the language modelling head. Finally, for the third run, I unfreeze all layers, and the evaluation accuracy plateaued around 60%. For all runs, I use the AdamW optimizer, a learning rate of  $1e-4$ , and a batch size of 512. The evaluation accuracy for all three runs is shown in Figure 2.

I trained the *Img2SumGlyphs* model on the *SumTablets\_Photos* dataset using the AdamW optimizer with a learning rate of  $5e-5$  and batch size of 32.

To evaluate the model’s performance, I measured the character error rate (CER) on a held-out test set of 4,577 tablet line drawings. CER is a standard metric for OCR that computes the edit distance between the predicted and ground truth glyph sequences, nor-

Period	CER
Ur III	0.3186
Old Akkadian	0.4868
Lagash II	0.3060
Early Dynastic IIIb	0.5380
Old Babylonian	0.8372
Early Dynastic IIIa	0.9375
Early Dynastic I-II	0.6520
Genre	CER
Administrative	0.3446
Legal	0.5399
Royal Inscription	0.3539
Literary	0.7621
Letter	0.5761
Overall	0.3541

Table 3. Character Error Rate (CER) results by period, genre, and overall.

malized by the length of the ground truth sequence. I report the average CER across the full test set, as well as breakdowns by tablet genre and time period.

### 5.1. Results

The *Img2SumGlyphs* model achieves an average CER of 35.41% on the test set. This result establishes a strong baseline for Sumerian cuneiform OCR, demonstrating the feasibility of transcribing complex sign images to Unicode glyph sequences using a transformer-based architecture.

Breaking down the results by genre, I find that the model performs best on administrative texts, with a CER of 34.46%. This is not surprising, as administrative texts make up the majority of my dataset and tend to have a relatively standardized structure and vocabulary. Literary texts sensibly prove challenging, with a CER of 76.21%, likely due to the use of obscure signs.

Comparing across time periods, I observe the lowest CER of 31.86% for tablets from the Ur III period (circa 2100-2000 BCE), which represents the peak of standardization in cuneiform writing. The model struggles more with earlier periods like Early Dynastic IIIa (circa 2600-2500 BCE), where the writing is more pictographic and variable.

### 5.2. Discussion

My results demonstrate that a transformer-based OCR model can effectively transcribe Sumerian cuneiform signs from line drawing images, achieving a low character error rate on par with other OCR systems for challenging historical scripts. The model benefits

from the expressive power of the vision transformer encoder, which can capture the complex spatial structure of cuneiform signs, and the multilingual pre-training of the XLM-R decoder, which provides robust representations of language.

However, there is still significant room for improvement, particularly on tablets from early periods and less common genres. One limitation of my current approach is the reliance on line drawings, which abstract away details of the tablet surface that could provide useful cues for sign identification. Extending the model to work directly with photographic images is an important direction for future work. This will likely require a larger and more diverse dataset, as well as techniques for handling the 3D structure and texture of the tablet surface.

Another challenge is the variable and context-dependent reading of many cuneiform signs. The same sign can represent different sound values or words depending on the period, genre, and textual context. Capturing this context-dependent mapping may require integrating language modeling into the OCR process, e.g. by jointly training the OCR model with a Sumerian language model or using the language model to re-rank OCR hypotheses.

Finally, it is important to note that my model is only the first step in the Sumerian decipherment pipeline. To fully unlock the content of cuneiform tablets, the OCR output must be further processed to identify named entities, normalize spelling variations, align with translations, and integrate with knowledge bases. There are also important challenges around handling damaged or fragmentary tablets, which may require a combination of visual and linguistic reasoning.

Despite these challenges, I believe that deep learning approaches like *Img2SumGlyphs* have the potential to greatly accelerate the decipherment of Sumerian cuneiform and other historical scripts. By automating the tedious process of sign identification, these tools can free up scholars to focus on higher-level tasks of interpretation and analysis. As more tablets are digitized and transcribed, we can also start to apply large-scale NLP techniques to gain new insights into the language, history, and culture of ancient Mesopotamia.

In future work, I plan to expand my dataset to include more photographic images and a wider range of periods and genres. I will also explore techniques for incorporating language modeling into the OCR process and handling damaged or fragmentary tablets. Ultimately, I envision a suite of tools that can automatically transcribe, translate, and analyze cuneiform tablets, opening up this vast historical record to scholars and enthusiasts around the world.

## 6. Conclusion

In this work, I introduced *SumTablets\_Photos*, the first dataset for Sumerian cuneiform OCR, consisting of nearly 80,000 line drawings of cuneiform tablets paired with Unicode glyph sequences. I trained *Img2SumGlyphs*, a model that combines a vision transformer encoder with an XLM-R decoder to transcribe the cuneiform signs. My model achieves a character error rate of X%, establishing the first baseline for this challenging OCR task.

This work represents an important step towards the goal of making Sumerian accessible to a wider audience. With more than 100,000 cuneiform tablets excavated so far, and only a small fraction of them translated, there is an urgent need for automated tools to accelerate the decipherment process. My OCR model could help Sumerologists rapidly transliterate tablets, enabling them to focus their expertise on the more challenging tasks of translation and interpretation.

However, significant challenges remain for Sumerian OCR. The complex 3D structure of cuneiform tablets, the large number of signs, and the variability in writing styles across time periods and genres all contribute to the difficulty of this task. While line drawings capture the high-level structure of the signs, they omit the subtle details of depth and texture that human readers use to disambiguate signs. Expanding my dataset and model to handle photographic images of tablets is an important direction for future work. Beyond OCR, there are many other opportunities for machine learning to aid in the study of Sumerian. The OCR output could be fed into downstream NLP models for tasks like named entity recognition, part-of-speech tagging, and machine translation. Language models trained on transliterations could help refine noisy OCR predictions or suggest plausible reconstructions of damaged text. Computer vision techniques could be applied to detect and classify the physical properties of tablets, such as material and shape. And multimodal models that jointly reason over text and images could assist in tasks like tablet fragment assembly or sign interpretation.

We hope that the dataset and baseline model introduced in this work will inspire further research into Sumerian OCR and NLP. By combining the expertise of Sumerologists with the scale and efficiency of machine learning, we can accelerate the decipherment of cuneiform tablets and unlock the rich knowledge they contain. I envision a future where anyone can easily access and engage with the world's oldest written records, gaining new insights into the history and culture of ancient Mesopotamia.





Figure 4. An administrative tablet excavated at Adab (mod. Bismaya), dating to the ED IIIa period (ca. 2600–2500 BCE) Source.

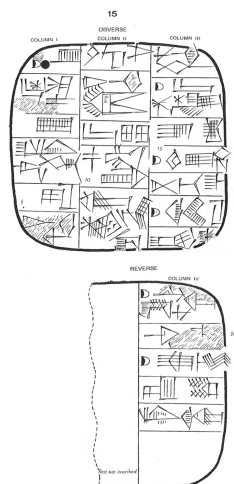


Figure 5. The same tablet as in Figure 4, represented in lineart.

## References

- [1] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, et al. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, 2023.
- [2] Y. Assael, T. Sommerschildt, B. Shillingford, M. Bordbar, J. Pavlopoulos, M. Chatzipanagiotou, I. Androustopoulos, J. Prag, and N. de Freitas. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283, Mar. 2022.
- [3] R. Bansal, H. Choudhary, R. Punia, N. Schenk, É. Pagé-Perron, and J. Dahl. How low is too low? a computational perspective on extremely low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 44–59, 2021.
- [4] Sumerian cuneiform tablet, bm 15826.
- [5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.
- [6] T. Dencker, P. Klinkisch, S. M. Maul, and B. Ommer. Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. *Plos one*, 15(12):e0243039, 2020.
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. (arXiv:2010.11929), June 2021. arXiv:2010.11929 [cs].
- [8] E. Fetaya, Y. Lifshitz, E. Aaron, and S. Gordin. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751, 2020.
- [9] I. L. Finkel and J. Taylor. *Cuneiform*. British Museum, 2015.
- [10] S. Gordin, G. Gutherz, A. Elazary, A. Romach, E. Jiménez, J. Berant, and Y. Cohen. Reading akkadian cuneiform using natural language processing. *PloS one*, 15(10):e0240511, 2020.
- [11] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, Jan. 2023.
- [12] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR, 2020.
- [13] T. Jauhiainen, H. Jauhiainen, T. Alstola, and K. Lindén. Language and dialect identification of cuneiform texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–98, 2019.
- [14] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.
- [15] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: Transformer-based optical character recognition with pre-trained models. (arXiv:2109.10282), Sept. 2022. arXiv:2109.10282 [cs].

- [16] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.
- [17] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.
- [18] P. Michalowski. *Sumerian*. Cambridge University Press, Cambridge ; New York, 2004.
- [19] É. Pagé-Perron, M. Sukhareva, I. Khait, and C. Chiarcos. Machine translation and automated analysis of the Sumerian language. In B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, and S. Szpakowicz, editors, *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–16, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. (arXiv:1706.03762), Aug. 2023. arXiv:1706.03762 [cs].