

# *SumTablets* 🏺: A Transliteration Dataset of Sumerian Tablets

**Cole Simmons**  
Stanford University  
coles@stanford.edu

**Richard Diehl Martinez**  
University of Cambridge  
rd654@cam.ac.uk

**Dan Jurafsky**  
Stanford University  
jurafsky@stanford.edu

## Abstract

Sumerian transliteration is a conventional system for representing a scholar’s interpretation of a tablet in the Latin script. Thanks to visionary digital Assyriology projects such as ETCSL, CDLI, and Oracc, a large number of Sumerian transliterations have been published online, and these data are well-structured for a variety of search and analysis tasks. However, the absence of a comprehensive, accessible dataset pairing transliterations with a digital representation of the tablet’s cuneiform glyphs has prevented the application of modern Natural Language Processing (NLP) methods to the task of Sumerian transliteration.

To address this gap, we present *SumTablets*, a dataset pairing Unicode representations of **91,606 Sumerian cuneiform tablets** (totaling **6,970,407 glyphs**) with the associated transliterations published by Oracc. We construct *SumTablets* by first preprocessing and standardizing the Oracc transliterations before mapping each reading back to the Unicode representation of the source glyph. Further, we retain parallel structural information (e.g., surfaces, newlines, broken segments) through the use of special tokens. We release *SumTablets* as a Hugging Face Dataset (CC BY 4.0) and open source data preparation code via GitHub.

Additionally, we leverage *SumTablets* to implement and evaluate two transliteration baselines: (1) weighted sampling from a glyph’s possible readings, and (2) fine-tuning an autoregressive language model. Our fine-tuned language model achieves an average transliteration character-level F-score (chrF) of 97.55, demonstrating the immediate potential of transformer-based transliteration models in allowing experts to rapidly verify generated transliterations rather than manually transliterating tablets one-by-one.



colesimmons/SumTablets (CC BY 4.0)



colesimmons/SumTablets



Figure 1: An administrative Sumerian cuneiform tablet from Shuruppak, dated to the Early Dynastic IIIa period (ca. 2500 BCE). (British Museum, 1896)

## 1 Introduction

Sumerian is the world’s earliest attested written language, marking the transition from prehistory into history as well as reflecting a rich written tradition spanning three thousand years. These texts are an invaluable resource in the study of ancient Near Eastern culture, politics, economics, and more.

During the latter half of the fourth millennium BCE, a sophisticated record-keeping system emerged in southern Mesopotamia, now known as proto-cuneiform (Selz, 2020). Over time this system evolved<sup>1</sup> to handle natural language. By about 2900 BCE this writing system, known as *cuneiform*, is concretely recognizable as encoding Sumerian.

Mesopotamian scribes originally devised the cuneiform script to write Sumerian. This script was later adapted to encode other languages throughout the Near East, such as Akkadian. To form glyphs, scribes would typically compose stylus impressions on a wet clay tablet<sup>2</sup>. Because

<sup>1</sup>There continues to be considerable ambiguity and disagreement about the extent to which evolution occurred gradually or was the result of a single inventor. For a more comprehensive treatment of the topic, see (Sprout, 2023).

<sup>2</sup>Although not all texts are clay or in the form of a tablet,

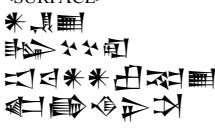

ID	Period	Genre	Glyphs (Inputs)	Transliteration (Targets)
Q001103	Early Dynastic IIIb	Royal Inscription	<SURFACE> 	<SURFACE> {d}en-lil <sub>2</sub> lugal kur-kur-ra ab-ba dingir-dingir-re <sub>2</sub> -ne-ke <sub>4</sub> inim gi-na-ni-ta ...

Table 1: A sample paired glyph–transliteration example from *SumTablets*, dating ca. 2600–2300 BCE.

cuneiform writing was impressed or inscribed on durable materials, texts have survived to the present in tremendous quantity (Finkel and Taylor, 2015). Uncovered during archaeological excavations of ancient cities beginning in the nineteenth century CE, these tablets had to be subsequently deciphered. Deciphering Sumerian, a language isolate, proved particularly challenging, and some periods and genres are still not completely understood.

Sumerian cuneiform glyphs are frequently polyvalent; that is, they have many possible readings (of no necessary semantic or phonetic relation) depending on the context. For instance,  can be read as *ka* “mouth,” *dug*<sub>4</sub> “to speak,” *kiri*<sub>3</sub> “nose,” *zuh* “to steal,” the syllable *ka*, and more. When reading a tablet, an Assyriologist must often consider various possibilities for each glyph to achieve a set of consistent readings. They represent their interpretation through the process of transliteration.

Transliteration is a modern, conventional system for representing Sumerian in the Latin alphabet. Conventions were established at various points in the modern, 150-year history of Sumerian decipherment and do not necessarily reflect the current understanding of Sumerian phonology or morphology. In transliterations, homophones are distinguished via subscripts; for instance, *e* and *e*<sub>2</sub> are homophonic—but semantically unrelated—readings of different glyphs. Additionally, hyphens are used to join nominal/verbal roots with affixes (Michalowski, 2004).

In 1996, the Electronic Text Corpus of Sumerian Literature (ETCSL) (Black et al., 2016) project began publishing transliterations online. This project became archival in 2006, soon followed by other projects such as the Cuneiform Digital Library Initiative (CDLI) (CDLI contributors, 2024) and the Open Richly Annotated Cuneiform Corpus (Oracc) (ORACC contributors, 2024). Thanks to

we follow Assyriological convention by referring to texts generically as tablets.

these and other projects, a large number of transliterations have been published online and their data made available for use with open licenses. Our work would not be possible without the decades of dedicated efforts by contributors to these projects.

Because Assyriologists are reading from either the physical text or an image, no digital representation of the original text’s glyphs is typically recorded. Today, most cuneiform glyphs have been added to Unicode<sup>3</sup>. However, easily accessible<sup>4</sup>, standardized datasets of paired Sumerian Unicode glyphs and transliterations remain limited, barring the development of transliteration models.

In this paper, we present the first large-scale, easily accessible dataset of **91,606 Sumerian tablets** as glyph–transliteration pairs, containing a total of **6,970,407 glyphs**. We additionally include IDs<sup>5</sup>, period, and genre metadata for each tablet to be used for results analysis.

Our dataset, *SumTablets*, is derived from a collection of publicly available Sumerian language resources, primarily the Electronic Pennsylvania Sumerian Dictionary (ePSD2) (Tinney et al., 2024) and the Oracc Sign List (OSL) (Veldhuis et al., 2024). These projects aggregate and index transliteration data from across Oracc, which shares data with CDLI and includes data from other current and former projects<sup>6</sup>.

Because of how they are formatted and because they do not include parallel Unicode glyph tablet representations, however, the data on Oracc are not immediately suited for glyph-to-transliteration tasks. We preprocess these data to clean and standardize them, converting structure-related anno-

<sup>3</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>4</sup>We define *easily accessible* as being easily utilized programmatically and requiring no or minimal Assyriological expertise to contribute to development of models based on these datasets.

<sup>5</sup>IDs are consistent with those in Oracc and CDLI.

<sup>6</sup>ePSD2 credits

tations into special tokens. Then, since a given reading maps back to only one glyph, we utilize Unicode–reading dictionaries provided by ePSD2 and OSL to convert each reading back into its source glyph.

We upload our dataset to Hugging Face ([HuggingFace Inc., 2024](#)), the largest and most widely utilized library for sharing datasets for machine learning tasks. We intend to use Hugging Face’s git-based version control to provide experiment reproducibility over time, with versions containing snapshots of the continuously updated Oracc data.

Our dataset, *SumTablets*, builds on previous open-source projects by:

1. **being the largest dataset of parallel glyph–transliteration examples.**
2. **standardizing the data available in Oracc**, optimizing formatting for the transliteration task while maintaining the morphosyntactic fidelity of the texts.
3. **vastly facilitating the use of this data in machine learning projects**, simplifying access via the common Hugging Face Datasets library.

Using our dataset, we develop and compare two baseline transliteration approaches. The first is a weighted dictionary mapping; for each glyph we sample one of the glyph’s possible readings according to its frequency. The second is a language model that we fine-tune for the glyph-to-transliteration task. As far as we are aware, we are the first to develop an automatic Sumerian transliteration model. Evaluated on a held-out test set, the dictionary-lookup approach obtains a character-level F-score (chrF) ([Popović, 2015](#)) of 61.22, while the fine-tuned model achieves a chrF score of 97.54.

Our goals in releasing this dataset are to facilitate the development of transliteration models and to demonstrate the potential of adapting large pretrained multilingual models for the task. We envision web-based tooling built on top of neural transliteration models helping Assyriologists to generate transliterations more quickly—allowing them to rapidly validate model outputs rather transliterating each tablet from scratch—and target review of potential errors in existing transliterations. Additionally, transliteration models serve as an essential step in eventually developing a complete Sumerian translation pipeline.

Finally, as a language isolate, Sumerian poses a unique syntactic challenge for cross-lingual models, and opens new avenues of research into the transfer of language understanding.

## 2 Related Work

To the best of our knowledge, our work represents the first to formulate Sumerian transliteration as an NLP task and to develop a transliteration model. However, prior works have utilized NLP techniques for other tasks in parsing and analyzing Sumerian cuneiform. The Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC) project ([Pagé-Perron et al., 2017](#)) sought to develop a pipeline for Sumerian annotation, translation, and information extraction, working primarily with Ur III transliterations. [Chiarcos et al.](#) expanded this data to include the Electronic Text Corpus of Sumerian Royal Inscriptions (ETCSRI) ([Zólyomi et al., 2019](#)). [Bansal et al.](#) then used MTAAC data in conjunction with CDLI and ETCSL data to train models for part-of-speech (POS) tagging, named entity recognition (NER), and translation, aiming primarily to build generalizable cross-lingual methods for performing these tasks on low-resource languages. The COMPASS ([Veldhuis, 2024](#)) also explores using cuneiform data for research tasks, such as reconstructing social graphs. Perhaps most similar to our work, [Gordin et al.](#) develop a neural network to automatically transliterate Akkadian from Unicode cuneiform glyphs.

Others have built datasets also representing tablets’ glyphs in Unicode. [Jauhiainen et al.](#) utilized Oracc dataset to build a dataset of 13,662 tablets for the task of language and dialect identification. More recently, [Chen et al.](#) used CDLI data to create CuneiML, a dataset of 38,947 tablets with photos, Unicode glyphs, transliterations, and metadata, also designed primarily for classification tasks. Both of these datasets include both Sumerian and Akkadian texts, whereas our dataset only includes monolingual Sumerian texts. Furthermore, our dataset is larger, designed specifically for the transliteration task, and is easily accessible through Hugging Face.

Outside of NLP, an exciting area of research is using computer vision methods to identify cuneiform signs from images ([Dencker et al., 2020](#)). Efforts in visual classification and transcription of cuneiform are enabled by projects that

have open-sourced high-quality 2D and 3D images of tablets (Dahl et al., 2019; Mara and Homburg, 2023). And beyond cuneiform, Assael et al. used deep learning methods to restore fragmented ancient texts in ancient Greek.

As Sumerian is a low-resource language, it is infeasible to train a transformer-based language model on Sumerian from scratch rather than adapting cross-lingual representations in existing models. Fortunately, the recent success of large cross-lingual NLP models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mT5 (Liu et al., 2020), and BLOOM (BigScience Workshop et al., 2023) have steadily raised the bar for zero- and few-shot cross-lingual performance on benchmarks such as XTREME (Hu et al., 2020) and MEGA (Ahuja et al., 2023). Recently, benchmarks to measure a model’s ability to perform NLP tasks in extremely low-resource and orthographically-diverse languages have emerged, such as IndicXNLI (Aggarwal et al., 2022) for low-resource Indian languages, and Sukhareva et al. who develop a POS tagging benchmark for Hittite, another cuneiform language. *SumTablets* marks the first benchmark for Sumerian neural machine transliteration.

### 3 Creating *SumTablets* 🍷

*SumTablets* is built upon the metadata and transliterations provided by ePSD2 via JSON files<sup>7</sup>. These transliterations were created or manually typed by scholars working in different projects around the world over decades of evolving knowledge of Sumerian vocabulary and grammar; they also contain extensive (but not useful for our purposes) embedded ASCII annotation. We begin by preprocessing the transliterations to normalize conventions, remove annotations, and convert formatting information into special tokens. Then, we use dictionaries built from ePSD2 and OSL resources to map each reading back to a Unicode representation of its source glyph. The result is a set of Unicode glyph–transliteration pairs with parallel formatting, allowing language models to most effectively learn the relationships between the two representations.

<sup>7</sup><https://oracc.museum.upenn.edu/epsd2/json>

### 3.1 Initial Data Cleaning

We first parse and type-check the ePSD2 JSON data using custom Pydantic<sup>8</sup> classes. The transliterations are structured in a recursive format called cdl (for the three node types: chunk, delimiter, and lemma) at the document level, which we navigate in order to reconstruct the transliteration as a single string with embedded formatting information.

We then remove annotations embedded in the transliterations. Many of these represent the editor’s interpretation beyond what is visible on the tablet; for instance, text enclosed in square brackets represents the editor’s belief of what was originally in a now-missing segment. While this information is academically useful, it can inject an undesirable bias when training transliteration models. Our goal is to best represent only what is on the tablet. We remove text enclosed in square brackets (broken) and single angle brackets (graphemes must be supplied for the sense but are not present), replacing the former with a ... special token to indicate breakage. For text enclosed in upper square brackets (partially visible) and double angle brackets (graphemes are present but must be excised for the sense), we remove the notation but retain the text. These examples are a few of many conventions used in the provided transliterations. For each type, we either remove the notation but retain the text, remove the notation and the text, or replace the notation and text with a special token (described in subsection 3.3).


The Oracc data are supplied with metadata that varies depending on the project in which a tablet was digitized. After performing an inner join on all of the data, we found the period and genre to be the most salient, universally-supplied metadata; because we provide the original Oracc IDs, removed fields can easily be reintegrated.


### 3.2 Mapping transliterations to glyphs

For each of the transliterations, we generate the associated glyphs in three steps:

1. First, we split each transliteration by spaces to get a list of words, which we then split further into individual glyph readings (i.e., morphemes).
2. Next, for each reading, we look up the corresponding glyph name. Each glyph in Sumerian has a conventional name that is an up-

<sup>8</sup><https://docs.pydantic.dev/latest/>

percase version of one of its readings; for instance, the glyph  is referred to as KA. Like most glyphs, it can be read a number of different ways (e.g., *ka*, *dug<sub>A</sub>*, *inim*). Importantly, these readings are readings only of KA and can be mapped back to it. If we cannot ascertain the glyph name, we replace the reading with <UNK>. Sign names are often used in place of a reading (to say that the reading is uncertain), in which case we replace the reading with <UNK> but will still use the corresponding Unicode. The first row of Table 2 shows the proportion of readings that we are able to map to glyph names.

- Finally, we convert each glyph name to the Unicode representation of that glyph name; for instance, we convert BU to . For the rare glyphs that are not represented in Unicode, we replace both the glyph and associated reading with <UNK> tokens. The bottom row of Table 2 shows the proportion of glyphs names that we able to map to Unicode.

To map from transliteration to glyph name and from glyph name to Unicode, we leverage ePSD2 and OSL.

Preprocessing Step	Success Rate
Readings → Glyph Name	6,724,498 (99.93%)
Glyph Name → Unicode	6,638,081 (99.96%)

Table 2: Preprocessing steps with associated amount of maintained glyphs in constructing *SumTablets*.

### 3.3 Extra-semantic tokens

In addition to the aforementioned preprocessing steps, we add the following special tokens to maintain structural information about each tablet in corresponding locations in the glyph and transliteration examples:

- **<SURFACE>** – The start of a surface. For a tablet, this may be the start of the obverse or reverse side. For other types of artifacts (like statues), the number of surfaces and their relationship to each other depends on the form.
- **\n** – A line break. These are important to retain because it is extremely rare that a word-form runs over to a subsequent line.
- **...** – Breakage. Ellipses on their own line indicate an indeterminate number of missing

lines, while ellipses on a line with text indicate an indeterminate number of missing glyphs.

- **<RULING>** – A horizontal line drawn by the scribe to separate sections of the tablet.
- **<COLUMN>** – The start of a new column of text. Not all tablets are formatted in columns.
- **<BLANK\_SPACE>** – The scribe left some amount of blank space before continuing on.

### 3.4 Metadata

As part of the dataset, we include additional metadata associated with each tablet: the time period each tablet dates from and the semantic genre of each tablet (e.g. administrative, legal). In total, we define 10 unique time periods and 14 genres (see Table 3).

Period	Train	Val	Test
Ur III	71,116	3,951	3,951
Old Akkadian	4,766	265	265
Early Dynastic IIIb	3,467	192	192
Old Babylonian	1,374	73	73
Lagash II	788	44	44
Early Dynastic IIIa	755	42	42
Early Dynastic I-II	77	4	4
Unknown	68	4	4
Neo-Assyrian	20	1	1
Neo-Babylonian	14	1	1
Middle Babylonian	7	0	0
<b>Total</b>	<b>82,452</b>	<b>4,577</b>	<b>4,577</b>

Genre	Train	Val	Test
Administrative	77,193	4,259	4,291
Royal Inscription	2,611	151	146
Literary	1,000	63	62
Letter	718	48	33
Legal	544	35	36
Unknown	269	14	7
Lexical	69	0	0
Liturgy	40	4	1
Math/Science	8	3	1
<b>Total</b>	<b>82,452</b>	<b>4,577</b>	<b>4,577</b>

Table 3: Composition of tablets by period and genre in *SumTablets*.

### 3.5 Data Partitions

For the purposes of developing automatic transliteration approaches, we split our corpus into train, validation, and test partitions using a 90%/5%/5% split. As an artifact both of what was produced as well as what sites have been excavated, there is a considerable imbalance in the number of examples between historical periods and genres. To ensure that we are training, validating, and testing evenly on how the language was used over time, we stratify the splits by period—Table 3 shows the number of examples in each by split. Because the genres of texts produced correlates strongly with period, stratifying by period results in a nearly equal split of genres, also shown in Table 3. Importantly, we removed the lexical texts before splitting, and then added them back to the train set after.<sup>9</sup>

## 4 Evaluating Transliteration Performance

The scale and standardization of *SumTablets* enables new methods to be applied to the task of Sumerian transliteration. In this section, we leverage our dataset to develop and compare two transliteration approaches: a straight-forward ‘dictionary baseline’ and a ‘neural baseline’. First, we define the transliteration task.

### 4.1 The Transliteration Task

We model transliteration as a sequence-to-sequence conversion task, where the input sequence is defined as glyphs and the output as a sequence of alpha-numeric characters, hyphens and white spaces. Table 1 illustrates example pairs of input (glyphs) and output (transliterations). As we model it, the transliteration task is more akin to a translation task, where each input sequence can be mapped to a large space of output sequences, rather than a token classification task. Given our framing of the transliteration task, we use character-level chrF score as the evaluation metric, defined as:

$$\text{chrF} = (1 + \beta)^2 \frac{\text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}} \quad (1)$$

where chrP and chrR stand in for character-level precision and recall scores. Throughout our

<sup>9</sup>Lexical texts are lists of words that were used in scribal training. We believe that it does not make sense to evaluate against them, but leave it up to the user to decide whether they provide productive noise during training.

analysis, we set  $\beta = 2$ , and use a character n-gram order of 6, as proposed by Popović. We compute the chrF score over the transliterated tokens for each tablet individually and then average these scores together over the dataset.

### 4.2 Dictionary Baseline

As part of previous transliteration efforts, Sumerian language experts have hand-crafted dictionaries that map a glyph to all possible readings of that glyph. We cross-analyze our dataset with the ePSD2 and OSL Sumerian dictionaries and find that the average number of different readings for a glyph, weighted by glyph frequency, is 22.17.

The availability of these dictionaries yields a simple automatic Sumerian transliteration approach: for each glyph in the test set, sample over its possible readings in proportion to their frequency<sup>10</sup>. This baseline results in an average chrF score of 61.22.

### 4.3 Neural Baseline

We explore whether the cross-lingual abilities of existing multilingual language models can be leveraged to solve the Sumerian transliteration task. Although Sumerian is a language isolate, it shares grammatical features with other modern languages: like Basque, it has ergative-absolutive alignment; like Turkish and Japanese, it is agglutinative; and like Korean, it is SOV (Michalowski, 2004). Therefore, the key to our approach is to leverage XLM-R (Conneau et al., 2020), a transformer language model pre-trained on over 100 languages.

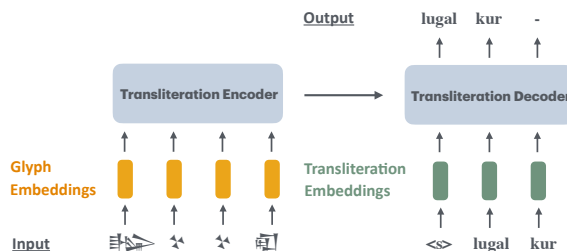


Figure 2: Illustration of the neural baseline model architecture. Inputs are read in as glyph tokens, while outputs are transliteration tokens.

The lack of tokenization support for Sumerian presents a first challenge in applying the XLM-R model to transliterating Sumerian. To deal with

<sup>10</sup>We recorded occurrence counts in the process of constructing the dataset.

Period			Genre		
Category	Dictionary	Neural	Category	Dictionary	Neural
Ur III	62.89	<b>98.46</b>	Administrative	63.15	<b>98.14</b>
Old Akkadian	64.52	<b>94.03</b>	Royal Inscription	54.58	<b>95.15</b>
Early Dynastic IIIb	62.51	<b>97.08</b>	Literary	37.73	<b>90.67</b>
Old Babylonian	37.70	<b>90.38</b>	Letter	47.43	<b>90.99</b>
Lagash II	58.55	<b>93.97</b>	Legal	56.19	<b>96.14</b>
Early Dynastic IIIa	67.85	<b>95.02</b>	Unknown	69.84	<b>97.58</b>
Early Dynastic I-II	73.72	<b>96.82</b>	Liturgy	55.92	<b>77.68</b>
Unknown	64.98	<b>89.87</b>	Math/Science	62.00	<b>95.12</b>
Neo-Assyrian	40.83	<b>89.79</b>			
Neo-Babylonian	42.47	<b>97.81</b>			
<b>Overall</b>	61.22	<b>97.54</b>		61.22	<b>97.54</b>

Table 4: Results by period and genre. Average chrF scores of transliterations generated in the dictionary baseline compared against those generated in the neural baseline.

this, we retrain the default SentencePiece tokenizer (Kudo and Richardson, 2018) used by the XLM-R model twice: once to build a ‘glyph tokenizer’ that is trained only on the Sumerian glyphs in *SumTablets*, and once to build a ‘transliteration tokenizer’ that is trained only on the corresponding Sumerian transliterations in *SumTablets*. The ‘glyph tokenizer’ has a vocab size of 632 glyph tokens and is used by the encoder model to generate ‘glyph embeddings’ from a string of Unicode-encoded glyphs. The ‘transliteration tokenizer’ has a vocab size of 1024 transliteration tokens and is used by the decoder model to output transliterations. The vocabularies of both the glyph and transliteration tokenizers include eleven special tokens, including the extra-semantic special tokens discussed in section 3.3.

We structure our transliteration model as a sequence-to-sequence (encoder-decoder) model. We initialize both the encoder and decoder separately with the pre-trained weights of an XLM-R model.

We train the model in three stages: First, we independently fine-tune the pretrained encoder model on the Unicode cuneiform glyphs using a masked language modeling task (MLM). This step yields a model with effective internal representations for the glyphs. Then, we integrate the decoder, training the full encoder-decoder model to take glyph sequences as input and auto-regressively predict target transliterations token-by-token. To stabilize the auto-regressive training of the joint encoder-decoder model, we decompose

this process two stages. We first freeze the encoder weights (only training the decoder) for one-third of the time that we train the joint encoder-decoder model. For the rest of training, we unfreeze the encoder weights and allow both the encoder and decoder to receive gradient updates. Figure 2 showcases the encoder–decoder model architecture. An added benefit of using both an encoder and a decoder is that the encoder can function independently from the decoder to predict missing or unknown glyphs, as illustrated in Figure 3.

Both the encoder and decoder are initialized with the pre-trained weights of a 279 million parameter XLM-R model<sup>11</sup>. We initially fine-tune the encoder on the MLM task for 50 epochs, with sequences lengths of 64 tokens, a learning rate of 5e-05, batch size of 2,048, and 200 warmup steps. We set the MLM masking probability to 0.10 and use the same 80-10-10 masking procedure as in Devlin et al.. Next, the encoder-decoder with frozen encoder weights is trained with a learning rate of 1e-04 for 2 epochs. Finally, we unfreeze the encoder weights and train the full encoder–decoder model with a learning rate of 5e-05 for a further 4 epochs. For both encoder–decoder learning procedures, we set the train batch size to 128 and the number of warmup steps to 100. All training used the AdamW optimizer (Loshchilov and Hutter, 2019) and was run on a single A100 SXM 80GB. For transliteration generation, we use beam search decoding with a beam size of 5.

<sup>11</sup>For a full description of the XLM-R model, refer to: <https://huggingface.co/FacebookAI/xlm-roberta-base>

Throughout our experiments, we set the maximum sequence length to 128. For tablets with more than 128 glyphs, we divide both the pre-tokenized glyphs and transliterations by newlines—these divisions align due to how we design *SumTablets* to preserve tablet structures. We then tokenize chunks of  $N$  lines, with  $N$  decreasing in size progressively from 16 down to 1, until the resulting chunk contains slightly less than 128 tokens. This segmentation ensures that all resultant chunks contain a maximum amount of tokens within the valid sequence length.

After processing the data into chunks of sequence length 128, we find that the dataset comprises 178,208 administrative examples and 23,282 non-administrative examples. To address the imbalance, we up-sample non-administrative examples by a factor of 5 for the initial two epochs of training and then reduce the up-sampling factor to 3 for the remaining epochs.

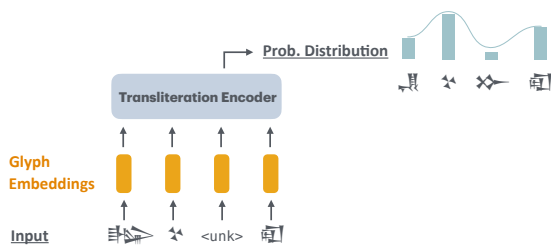


Figure 3: The encoder model can produce a probability distribution over possible glyphs that can replace an <UNK> token. This is because the encoder is trained using an MLM objective.

Our encoder-decoder model achieves an average character-level chrF score of 97.54 on unseen test data, setting, to the best of our knowledge, a new state-of-the-art benchmark performance on the transliteration task. We report results for both baselines and across all time periods and genres in Table 4. Our work demonstrates the capability of large multilingual models to model and transliterate Sumerian, despite the highly fragmented nature of these extant texts and the language being both low-resource and an isolate.

#### 4.4 Analysis

We derive several key takeaways from our results.

**The genre of the texts impacts transliteration performance.** The difference in transliteration performance across genres that we observe in Table 4 is intuitive given the nature of the underlying

data. Because the training data is dominated by administrative examples, it is natural that that would be the best performing category. These texts also tend to be relatively formulaic. Liturgical, letter, and literary texts, on the other hand, have a different style, form, and vocabulary from the rest of the corpus. These genres (liturgical in particular) are also some of the most challenging for experts. Genre also affects performance insofar as for most genres there are so few examples on which to train or evaluate.

**Inconsistent transliteration conventions muddy performance.** Some of the different readings for a glyph stem not from a tangible semantic difference but from phonetic or aesthetic disagreement. For instance, “saj” and “sag” represent the same thing, but Assyriologists have a preference in how they represent the nasal ‘g.’ A lack of standardization on matters like this fragments the patterns in which models observe a reading occurring.

**It is difficult to predict phonemes in names.** Manual error analysis showed that some errors occurred when selecting a reading that serves as part of a name (playing a phonetic role). Our neural baseline model would often predict a valid reading for a glyph, but a different one than in the true transliteration. Future work will incorporate expert evaluation to determine whether these predictions are any more or less plausible than those in the original transliteration.

## 5 Limitations

We note that our work has some limitations, both in terms of the *SumTablets* dataset and the transliteration model.

### 5.1 Dataset Limitations

Administrative documents have an outsized representation in the train, validation, and test data. This dataset imbalance is a natural by-product of the category of documents produced by Mesopotamian peoples and is an unavoidable consequence of working with Sumerian texts. Although we chose to oversample non-administrative tablets in the train set by a factor of 5 during training of our model, we leave the choice of how to best handle this imbalance to the consumer.

While the set of Unicode cuneiform glyphs is largely complete, there are still glyphs that are not represented in this set, particularly some com-



plex compound glyphs. We currently convert these glyphs and their corresponding readings into <UNK> tokens, but future work could incorporate unique identifiers for these glyphs as a placeholder until they are added to the Unicode standard.

Finally, there is considerable orthographic variation in glyphs over time, and representing these in Unicode flattens these (potentially meaningful) variations into a single, universal representation.

## 5.2 Model Limitations

In this paper, we train an XLM-R model on *SumTablets* as a fully supervised neural baseline for Sumerian glyph transliteration. We give our model access to the entire training set to explore the limit of a pre-trained cross-lingual model to perform this novel task. Our work, however, does not study the zero- and few-shot abilities of cross-lingual models, which is typically of more interest when evaluating a model’s cross-lingual abilities. Nor do we study the performance of a model trained from scratch on our dataset. We encourage future work to use *SumTablets* as a few- and zero-shot cross-lingual benchmark task to evaluate how a multilingual model’s language understanding transfers to the Sumerian language.

Moreover, we recognize that the dictionary baseline that we implement is very simple, and that a better point of comparison would be an N-gram model.

## 6 Conclusion

We introduce *SumTablets*, the first collection of paired glyph-transliterations extracted from 91,606 Sumerian tablets. Our dataset provides a resource for experts and non-experts alike to contribute to the development of transliteration models. We define the transliteration task, evaluation method, and establish a baseline performance so that future results may be compared. We also demonstrate that—despite Sumerian’s status as a low-resource language and language isolate—large pretrained multilingual language models can be adapted to perform the sequence-to-sequence task of transliterating a sequence of Unicode cuneiform glyphs with remarkable accuracy.

With such an abundance of extant texts and so few specialists capable of reading them, we believe transliteration models will enable Assyriologists to spend less time on tedious, from-scratch transliteration and more time on research and translation.

## Acknowledgments

We would like to thank Niek Veldhuis for his assistance as we explored past digital and computational Assyriological work.

First and foremost we thank the relentless dedication of contributors to the countless open-source cuneiform and Sumerian language resources. We would also like to thank Niek Veldh, Sabri Eyuboglu and Geoff Angus for their many suggestions and guidance in creating SumTablets. Richard Diehl Martinez is supported by the Gates Cambridge Trust (grant OPP1144 from the Bill & Melinda Gates Foundation).

## References

- Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. [IndicXNLI: Evaluating Multilingual Inference for Indian Languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. [MEGA: Multilingual Evaluation of Generative AI](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. [Restoring and attributing ancient texts using deep neural networks](#). *Nature*, 603(7900):280–283.
- Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Émilie Pagé-Perron, and Jacob Dahl. 2021. [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, Online. Association for Computational Linguistics.
- BigScience Workshop, Teven Le Scao, Angela Yu, Victor Sanh, Albert Webson, Mariama Barbu, Xingxing Yap, Cécile Gessler, Sonja Lovrić, Niklas Muenighoff, Stella Biderman, et al. 2023. [BLOOM: A 176B-Parameter Open-Access Multilingual Language Model](#). *Preprint*, arXiv:2211.05100.
- J.A. Black, G. Cunningham, J. Ebeling, E. Flückiger-Hawker, E. Robson, J. Taylor, and G. Zólyomi.

2016. Electronic Text Corpus of Sumerian Literature. <https://etcsl.orinst.ox.ac.uk/>.
- British Museum. 1896. [Tablet BM 15826](#). Clay tablet; Pre-Sargonic. Sumerian account of silver and other commodities; square tablet, 8 col.
- CDLI contributors. 2024. Cuneiform Digital Library Initiative. <https://cdli.mpiwg-berlin.mpg.de/>.
- Danlu Chen, Aditi Agarwal, Taylor Berg-Kirkpatrick, and Jacobo Myerston. 2023. [CuneiML: A Cuneiform Dataset for Machine Learning](#). *Journal of Open Humanities Data*, 9(1).
- Christian Chiarcos, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 2018. Towards a Linked Open Data Edition of Sumerian Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jacob L. Dahl, Hendrik Hameeuw, and Klaus Wagnonsonner. 2019. [Looking both forward and back: imaging cuneiform](#). *Cuneiform Digital Library Preprints*, 2019(14.0). [Online; accessed 2024-07-04].
- Tobias Dencker, Pablo Klinkisch, Stefan M. Maul, and Björn Ommer. 2020. [Deep learning of cuneiform sign detection with weak supervision using transliteration alignment](#). *PLOS ONE*, 15(12):e0243039.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Irving L. Finkel and Jonathan Taylor. 2015. *Cuneiform*. British Museum.
- Shai Gordin, Gai Gutherz, Ariel Elazary, Avital Romach, Enrique Jiménez, Jonathan Berant, and Yoram Cohen. 2020. [Reading Akkadian cuneiform using natural language processing](#). *PLOS ONE*, 15(10):e0240511.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- HuggingFace Inc. 2024. Huggingface datasets. <https://huggingface.co/docs/datasets/index>.
- Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019. [Language and Dialect Identification of Cuneiform Texts](#). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–98, Ann Arbor, Michigan. Association for Computational Linguistics.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Hubert Mara and Timo Homburg. 2023. [MaiCuBeDa Hilprecht - Mainz Cuneiform Benchmark Dataset for the Hilprecht Collection](#).
- Piotr Michalowski. 2004. Sumerian. In Roger D. Woodard, editor, *The Cambridge Encyclopedia of the World's Ancient Languages*. Cambridge University Press, Cambridge ; New York.
- ORACC contributors. 2024. Openly Richly Annotated Cuneiform Corpus. <https://oracc.museum.upenn.edu/>.
- Émilie Pagé-Perron, Maria Sukhareva, Ilya Khait, and Christian Chiarcos. 2017. [Machine Translation and Automated Analysis of the Sumerian Language](#). In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–16, Vancouver, Canada. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: Character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Gebhard J. Selz. 2020. [The Uruk Phenomenon](#). In Karen Radner, Nadine Moeller, and D. T. Potts, editors, *The Oxford History of the Ancient Near East: Volume I: From the Beginnings to Old Kingdom Egypt and the Dynasty of Akkad*, pages 163–244. Oxford University Press.

- Richard Sproat. 2023. *Symbols: An Evolutionary History from the Stone Age to the Future*. Springer Nature Switzerland, Cham.
- Maria Sukhareva, Francesco Fuscagni, Johannes Daxenberger, Susanne Görke, Doris Prechel, and Iryna Gurevych. 2017. Distantly Supervised POS Tagging of Low-Resource Languages under Extreme Data Sparsity: The Case of Hittite. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 95–104, Vancouver, Canada. Association for Computational Linguistics.
- Steve Tinney, The Pennsylvania Sumerian Dictionary Project, and ePSD2 contributors. 2024. Electronic Pennsylvania Sumerian Dictionary. <https://oracc.museum.upenn.edu/epsd2/index.html>.
- Niek Veldhuis. 2024. [niekveldhuis/compass](https://github.com/niekveldhuis/compass). <https://github.com/niekveldhuis/compass>.
- Niek Veldhuis, Steve Tinney, and Robin Leroy. 2024. Oracc Sign List. <https://oracc.museum.upenn.edu/osl>.
- Gábor Zólyomi, Bálint Tanos, and Szilvia Sövegjártó. 2019. The Electronic Text Corpus of Sumerian Royal Inscriptions. <https://oracc.museum.upenn.edu/etcsri/index.html>.