AI Project Proposal

Luca Martucci, Michael Menghi, Samuele Straccialini, Francesco Vinciguerra

April 2025

For generations, the task of deciphering ancient languages has been left to researchers, who meticulously transcribed symbols by hand and translated them into modern languages. Recently, this process has been accelerated by AI-powered tools. One notable example is Google's Fabricius, which enables the recognition and translation of ancient Egyptian hieroglyphics, and has brought new attention to the potential of machine learning in historical linguistics. While several tools for automatic decipherment of hieroglyphics from ancient Egyptian artifacts have been developed, other pictogram-based languages have been less impacted by new technologies, such as Sumerian cuneiform.

Our project investigates a multi-stage pipeline, aiming to:

1. Translate transliterated cuneiform into English using Transformers;
2. Recognize individual signs from tablet images via CNN-based image classification;
3. Transcribe entire artifact images through image segmentation.

We are aware that the full pipeline is ambitious, so we will adopt an incremental approach, initially focusing on the translation task while progressively addressing the others as time and feasibility allow.

We will rely on the following datasets:

- `https://huggingface.co/datasets/colesimmons/SumTablets_English` and

  `https://huggingface.co/datasets/colesimmons/SumTablets_English-augmented`

  for the translation task. They contain the transliteration and English translation of artifacts.
- `https://huggingface.co/datasets/colesimmons/SumTablets` and

  `https://huggingface.co/datasets/colesimmons/SumTablets_Photos`

  for the transcription tasks. They contain the glyphs' transcription and transliteration of artifacts, together with images of photos and *"lineart"* (researchers' reproductions).
- We may need to generate synthetic data for image recognition or segmentation, as we do not have access to isolated glyph images. This can be done using Unicode characters corresponding to Sumerian pictograms, available in the *NotoSansCuneiform* font.

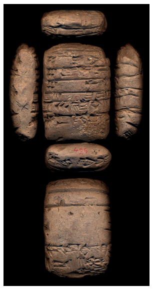

Figure 1: Sample artifact

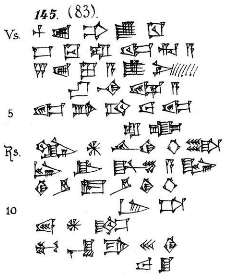

Figure 2: Sample *lineart*